DOCUMENT RESUME

ED 282 415 FL 016 706

AUTHOR	van Weeren, J., Ed.
TITLE	Practice and Problems in Language Testing 5. Non-Classical Test Theory; Final Examinations in Secondary Schools. Papers Presented at the International Language Testing Symposium (5th, Arnhem, Netherlands, March 25-26, 1982).
INSTITUTION	Centraal Inst. voor Toetsonwikkeling, Arnhem (Netherlands).
REPORT NO	30
PUB DATE	May 83
NOTE	108p.; For individual papers, see FL 016 707-715.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Adolescents; Comparative Analysis; Comparative Education; Difficulty Level; Evaluation Criteria; Evaluation Methods; Foreign Countries; Language Skills; *Language Tests; *Listening Comprehension; *Measurement Techniques; Modern Languages; Multiple Choice Tests; *Reading Comprehension; Reading Tests; Secondary Education; *Test Format; Testing; Testing Problems; Test Theory; Test Validity
IDENTIFIERS	*Comprehensive Examinations; England; Netherlands; Northern Ireland; *Rasch Model; Sweden; Wales

ABSTRACT

Presented in this symposium reader are nine papers, four of which deal with the theory and impact of the Rasch model on language testing and five of which discuss final examinations in secondary schools in both general and specific terms. The papers are: "Introduction to Rasch Measurement: Some Implications for Language Testing" (J. J. M. Theunissen); "Focusing in on a Latent Trait: An Attempt at Construct Validation by Means of the Rasch Model" (John H. A. L. de Jong); "Rasch and Reading" (Digna M. M. Samson); "Objective Based Tests for Reading Comprehension" (Willem M. M. van Roosmalen); "The Swedish System of Final Examinations in Modern Foreign Languages" (Torsten Lindblad); "Language Testing at About Age 16: Current Developments in England, Wales and Northern Ireland" (Robert H. Osborne); "Final Language Examinations in Secondary Education: The Dutch System" (Trude A. Maas); "Backwash Effect of Multiple Choice Language Tests: Myth or Reality?" (Hildo Wesdorp); and "Response Models for Examinations with Open-ended Questions" (Henk Kreeft, Piet Sanders). (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED282415

Practice and Problems in Language Testing 5

non-classical test-theory
final examinations in secondary schools

J. van Weeren (ed.)

Cito

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. van Weeren

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL016706

2

Algemene publikatie Nr 30

Practice and Problems in Language Testing 5

non-classical test-theory
final examinations in secondary schools

J. van Weeren (ed.)

BEST COPY AVAILABLE

Cito
Arnhem, May 1983

Preface

This reader includes the full text of papers, presented at the Fifth International Language Testing Symposium. The Symposium took place in Arnhem, The Netherlands on 25 and 26 March 1982. It was hosted by the Dutch National Institute for Educational Measurement (CITO) and organized in collaboration with the Interuniversitäre Sprachtestgruppe (IUS).

Selecting a main theme for the Symposium, the organizing committee opted for two topic areas: non-classical test-theory and final examinations in secondary schools. With the first rather technical topic a tradition of IUS, established at former conferences, was continued. This time a short introduction was given into Rasch measurement as the favourite exponent of non-classical test-theory. The next three papers demonstrate the impact of the Rasch-model on language testing by dealing with practical applications of this theory on: construct validation of a listening test as well as of a reading comprehension test in a foreign language. The first paper describes an investigation with native speakers used as a criterion; in the second paper different methods of measuring reading comprehension are compared with each other. The third paper describes the development of reading comprehension tests in the mother tongue which are homogeneous and specific for different ability levels.

The second main topic, final examinations in secondary schools, was discussed in both a general and a more specific way. Three speakers explained how quality control in education is set up by the central governments in Sweden, Great Britain and the Netherlands. Although these papers contain several interesting details on language testing, more specific problems are dealt with in the last two papers. The first one presents empirical evidence concerning the problem of 'backwash' effects and deals with the question to what extent they can be measured. The second paper deals with the interjudge-reliability of open-ended questions. Especially in the context of language testing open-ended questions often seem to be favoured, but how far can interjudge-reliability be guaranteed by providing elaborate lists of model responses?

We wish to thank all the authors of these papers for their participation in the Symposium as well as for allowing them to be published. We especially thank Marion Feddema for her co-ordination of both production and handling of this fine booklet.

J. van Weeren
Arnhem, May 1983

Contents

Preface 3

- Introduction to Rasch Measurement:
  Some Implications for Language Testing,
  Phiel J.J. M. Theunissen 7

- Focusing in on a Latent Trait:
  An Attempt at Construct Validation by means of the
  Rasch Model, John H.A.L. de Jong 11

- Rasch and Reading, Digna M.M. Samson 37

- Objective Based Tests for Reading Comprehension,
  Willem M.M. van Roosmalen 45

- The Swedish System of Final Examinations in Modern
  Foreign Languages, Torsten Lindblad 51

- Language Testing at about Age 16:
  Current Developments in England, Wales and Northern
  Ireland, Robert H. Osborne 67

- Final Language Examinations in Secondary Education:
  The Dutch System, Trude A. Maas 77

- Backwash Effects of Multiple Choice Language Tests:
  Myth or Reality? Hildo Wesdorp 85

- Response Models for Examinations with Open-ended
  Questions, Henk Kreeft/Piet Sanders 105

- Contributors 119

In stark contrast with the well-known 'boy-meets-girl' situation, surprisingly little is known about the 'pupil-meets-test' situation. As test developers, we know at least one thing for sure, that is, we have to measure the pupils ability or performance, and one way to do this is by using the Rasch-model. In this one sentence I have used two concepts, the understanding of which is important to an understanding of test-theory. These are the concepts of 1. measurement and 2. model.

What is measurement? A very simple definition is: 'measurement is the assignment of numbers to object people according to certain rules.' Examples of rules are the ways in which we measure length (e.g. using an object calibrated with a standard in Paris) or temperature in centrigrades. We find it then completely 'normal' that we can say things like: "This stick is twice as long as that stick, or 10° C is equivalent to 50° F ($F = \frac{9}{5} C + 32$)." Unfortunately, this level of sophistication is rarely reached outside the physical sciences. The meaningfulness of the numerical statements we make depends on the mathematical structure of the number system we use to measure. If we use only the ordinal properties of numbers, we can only make ordinal statements, as is basically the case in a lot of standard practical use of test results.
Unless we know what we are doing when we measure pupils' ability, there will never be any real progress in this field.

Any systematic approach towards the assessment of pupils' ability implies the use of a model. The more explicit the model is, the better we know what we are doing and what the consequences are. What is a model? Without going into detail, we can say that a model is an abstract picture of a 'chunk' of reality in which relevant factors are defined and their mutual relationships are specified. A model implies the use of basic assumptions, on which the model is based. A model can have few assumptions, in which case it is usually called a weak model, or may have more assumptions, and is thus called a strong model. A weak model has relatively little explanatory power, but is less easily rejected by the emperical data, while a strong model has relatively large explanatory power, but is more easily rejected by the data. Models can also differ on the following point: some models are formulated in terms that presuppose that measurement has already taken place, while in other models measurement is a result of the theory.

The Rasch-model is a strong model where measurement is a bonus of the model itself. The Rasch-model is based on four basic ideas:

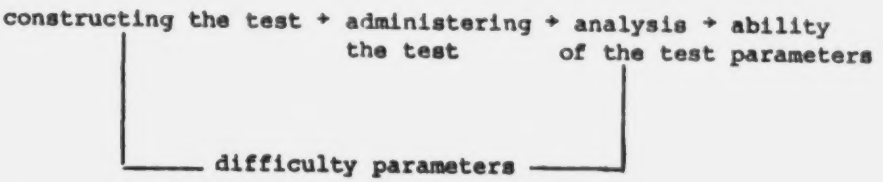
1 separation of observable behaviour and an underlying latent trait, responsible for this behaviour
2 observed behaviour is essentially probabilistic in character
3 Local stochastic independance. In terms of testing this means that the correlations between items are only caused by their relationship with the latent trait (a homogeneous test) and that in the absence of the trait, or when the trait is kept constant at a particular level, the correlation disappears.
4 Measurement of the latent trait requires the repeated assessment of indices of this trait, not necessarily the same index on each occasion.

The Rasch-model specifies that the performance of the pupil is completely determined by two variables, i.e.:
1 his ability and
2 the difficulty of the item
with the exclusion of other factors. So it excludes e.g. the time-factor, guessing behaviour and the influence of the discriminative power of the items (this is the degree of sensitivity in distinguishing between higher and lower ability). So the only things that can vary in the Rasch-model are the ability and the difficulty. We call these the parameters of the model, with symbols θ and σ. Since we take a probabilistic off round, any numerical result of the expression in which θ and σ play a role will lie in the interval $[0,1]$. It is perfectly reasonable to assume that the chance of a correct performance increases with an increase in θ and decreases with an increase in σ. Even more simply formulated one could say that the chance of correct performance increases, while $\theta - \sigma$ increases. One of the most simple statistical functions that can be used in the above situation is the logistic function.

$$P\,ij = P(+/\theta) = \frac{e^{(\theta j - \sigma i)}}{1 + e^{(\theta j - \sigma i)}}$$

The reasons for choosing this function and no other are too technical to mention here.

What are the noticeable characteristics of the Rasch-model?
1 The following sequence is appropriate to the Rasch-model:



We see here that we have a theory (of behaviour) before we have measurement. This is in contrast with classical test-theory, which assumes measurement and only then starts as a model ($X = T + E$). The origin of the measures is independent of the classical test-theory and in practice is rarely, if ever, rationalized, so that we very often do not even know the level of measurement.

2 The Rasch-model is geared towards the item-level, instead of the test-level, thereby using only qualitative information on item-level, that is, whether the item has been correctly constructed or not. Distilling quantitative information from qualitative data requires explicit theory that can be tested and, if necessary, rejected. In the Rasch-model, by quantifying the parameters, we 'explain' qualitative information. Classical test-theory uses a 'pseudo-quantitative' test-score as point of departure. The scale of measurement (the same for θ and σ) follows from the theory.

3 The Rasch-model does not try to explain in a deterministic manner why a certain answer to an item is given, but tries to explain the chance of a correct response.

4 The Rasch-model has the property of 'specific objectivity'; it is 'population-independent'. This means that comparison of ability-parameters is not dependent on a particular set of items being used and also that item-parameters can be compared, independent of the particular pupils being used to calibrate the items.

This property is common in physical measurement and is a very important property. We take it for granted, for example, that in measuring the length of a pole it is not important which yard-stick is used, neither that the outcome of the measurement is dependent on what sort of pole is being measured. It has been proven that the Rasch-model is the only model with this property.

This is all in clear contrast with classical test-theory, where item and test characteristics are really population-dependent. Classic measures like the p-values of the items and the reliability (KR20) of the test depend on the population from which the test-data were obtained

$(X = T + E)$

$(\text{Var } X = \text{var } T + \text{var } E)$

$$\text{Reli} = \frac{\text{var } T}{\text{var } X} = \frac{\text{var } T}{\text{var } T + \text{var } E} \qquad \frac{1}{1+1} = .5 \qquad \frac{3}{3+1} = .75$$

population-dependent

increase in var T in a population increases Reli, even if var E remains constant (= accuracy of measurement is constant)

5 The Rasch-model permits a clearer insight into the accuracy of ones' measurements. The Rasch-model allows a differential assessment of the accuracy of measurement, by computing the information-functions (-curves) for items and test.

This has several advantages:

a The concept of parallelism of test becomes much clearer. Two tests are defined as parallel if they measure the same ability and have the same test-information function. In this way, two or more seemingly very different tests can be made parallel.

b Using the test-information function in conjunction with the distribution of true ability of the pupils enables one to make statements about the applicability and usefulness of the tests.

Model-fit and the Rasch-model

In order to give a rough idea, one could say that, in very general terms, the assessment of model-fit could be done in the following way:

a collect observations
b estimate parameters
c use the model to make predictions about what data ought to be expected if the model were true
d find a measure to express the discrepancy between observation and expectation.

Assessing fit is still a problem. It is not really known how sensitive the existing measures of fit are to violations of what assumption(s).

Extensions of the Rasch-model

E.g.: Linear logistic model: The σ parameter can have a linear composition, expressing a theory about the difficulty of items.

An example of applications

Suppose we have a large number of items all measuring the same ability and calibrated according to the Rasch-model in a variety of different situations. It then becomes possible to recalibrate all items in such a manner that all subsequent values refer to a common scale. For a pair of tests, this could be done by the anchor item technique. However, as the number of items and their range of difficulty grows and grows, one has to think of using a network of linking tests, finally resulting in what is known as an itembank. If successfully constructed, one has assembled a very large number of items on a common scale, with a range of difficulty too large for any one single person to cope with at a given moment. From the bank, one can then generate, according to one's own specification, a large number of different tests, tailor-made, so to say. One can make easy tests, for more accurate assessment of low level ability, or difficult tests. One can make long tests and short tests, one can make tests with a wide range of difficulty or a narrow range, and so on. One can compare locally made items (e.g. on the school level) with nationally used items (e.g. Cito items), if both are in the bank. Obviously, an itembank can be used to assess development and growth, e.g. as a result of curriculum or other more sophisticated changes. A fancier application is a form of criterion-referencing whereby mastering a criterion is scored as if it were an item (1 = succes, 0 = failure) and introduced into the bank, together with other 'normal' items. Such a criterion could be passing or failing an exam, or admission, for example, into a professional training programme. The itembank can then be used to measure a person's approach to or departure from the criteria.

Focusing in on a Latent Trait: An Attempt at Construct Validation by means of the Rasch Model,

John H.A.L. de Jong

The Rasch model for test analysis is a so-called latent trait model. In a model of this kind, a relationship is specified between observable test performance and the unobservable traits or abilities assumed to underlie performance on the test.
In most cases the test constructor has no clue as to whether the latent traits postulated by the model are indeed the abilities he wants to measure. Try-out sessions of a listening comprehension test on native speakers and on foreign language learners have yielded data possibly offering a means to identify the ability which, according to Rasch test analysis, underlies the main trait measured by the test.

Introduction

As the subtitle of my paper indicates I will try to prove construct validity of a test by means of the Rasch model. The test I will deal with is a foreign language listening comprehension test, which has recently been developed at our Institute for Educational Measurement (Cito). I will start with a brief description of the test. Secondly I will define the ability, which we propose to measure with the test, which is in fact the theoretical construct. In the body of the paper I will offer evidence for a relationship between the construct as defined and the measure of fit of items according to the Rasch model. Finally I will illustrate the assumptions made in this paper with examples from the actual test.
After Phiel Theunissen's paper (this volume) I feel free to assume a certain knowledge of the Rasch model but I will explain concepts of particular importance for the evidence which I hope to give.

The test itself

It is important to make the preliminary remark that the test was developed as a research project: to investigate techniques of testing foreign language listening comprehension, techniques which were new to us. Obviously such a test will contain a larger number of items of poor quality than any of our tests constructed for actual use in final examinations.

The test uses life recordings taken from different radio programs, cut into rather short samples of spoken language (about 25 seconds each). The language to be tested is English. Testees listen to the tape, hear each sample only once and have to respond to a multiple choice question with two options printed in a test booklet. They have a 10 second pause in between samples to decide on their choice.

Two types of items were used:

A True-false items
Testees have to decide whether the statement in the test booklet is in accordance with what has been said on the tape.

B Modified cloze items with two options
Words to be deleted from the text were not randomly selected but were chosen for their semantic relevance. In each sample a word or group of words was cut out from the tape and replaced by an electronical sound. Testees were to decide which of the two options presented in their test booklet could be used to restore the text. Thus the typical problem with cloze items - acceptable word or exact word scoring - could be avoided and no productive skill was tested.

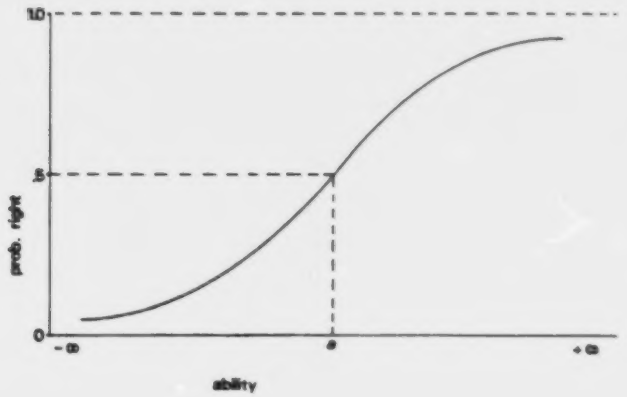
The test was tried out on two groups:
- a representative sample of 575 Dutch pupils taken from the target population (pupils preparing for final Secondary School examinations after 6 years of English as a foreign language)
- a group of 30 native speakers of comparable age and educational background

The construct

For construct validation it is necessary to determine the concept which accounts for performance on the test (American Psychological Association, 1974; Cronbach and Meehl, 1955). In a foreign language listening comprehension test we hope that difference in raw scores on the test can be explained by difference in ability to understand the spoken language in question.
What is the ability to understand a spoken language? Without going into the discussion on the Unitary Competence Hypothesis versus the belief that linguistic competence can be broken down into a number of totally distinct factors or Vollmer's suggestion of a hierarchical model (Vollmer and Sang, 1981; Vollmer, 1981) one can safely assume that native speakers of a language do possess this ability.
Still we all know that native speakers differ in their ability to follow the spoken language according to their ability to cope with the language material at the conceptual level.
Obviously, defining foreign language listening comprehension as 'the ability native speakers demonstrate in understanding spoken samples from their native tongue' is too comprehensive. To rule out non-linguistic factors the concept should be narrowed down to: 'The ability to understand the foreign language at the level of native speakers of comparable age and educational background'. A group of native speakers thus defined will have to do extremely well on the items in the test and in any case, they will have to do no less than the most able listeners in the target population of non-natives. Furthermore no significant variance in native speaker scores on the test is to be expected. I will not go into this point here but small variance in native speaker scores on the test has been reported (de Jong en van den Nieuwenhof, 1982).

In terms of the Rasch model: the model postulates unidimensionality of the ability or underlying trait. Because we expect native speakers to show greater ability in listening comprehension, they should have higher probability of getting the right answer on each item, if the item requires listening ability.

Figure 1: Item characteristic curve for item i



Now please consider figure 1 which represents the item characteristic curve for an imaginary item i. Different levels of ability (x-axis) are plotted against the chance of getting item i right (y-axis). A person of ability 'a' e.g. has exactly .50 chance to get the item right. A chance of 1.0 is of course the highest possible chance and would require ability at + ∞. In other words a chance of 1.0 is the upper limit, the ceiling for any person doing item i.

Figure 2: Item characteristic curve for item i

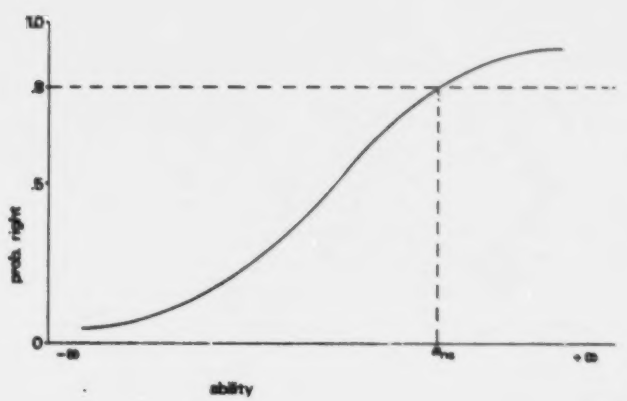
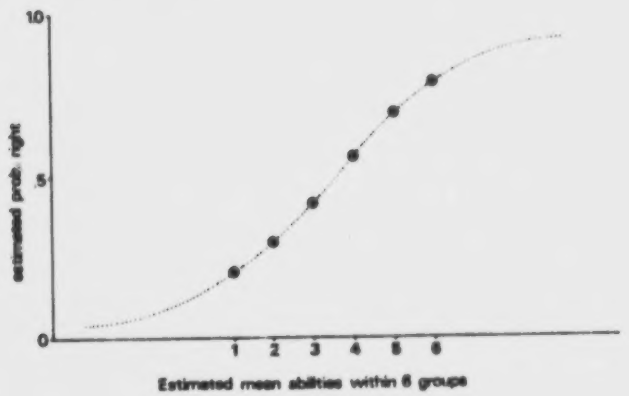


Figure 2 represents this relation between ability and probability of getting item i right again. On the y-axis a dotted line at .8 represents the observed proportion of right answers in a group of native speakers defined as before. From their observed responses we can estimate the

mean ability of this group of native speakers: a_{ns}.
And because we expect non-natives to show less ability in solving the item than native speakers - their ability level on the x-axis should be to the left of that of native speakers (a_{ns}) - we expect them to have less chance in getting the item right. This implies that for the target group of non-native speakers the upper limit of their chance to get the item right is set by the native speaker scores on that item: for this item .8 (and not 1.0). If they score higher than the native speakers then the ability required for the item is not the ability we set out to measure.

For this research Rasch analysis was done by computer with the program CALFIT (Wright and Mead, 1975). The unconditional maximum likelihood procedure (UCON) of this program was used to estimate ability and difficulty parameters (Wright and Stone, 1979).
The program divides testees into six groups of roughly the same size according to level of ability estimated from their performance on the test.
The program then calculates, for every item, the probability of success on that item for the six groups, and expresses the probability as an expected number of right answers in each group.

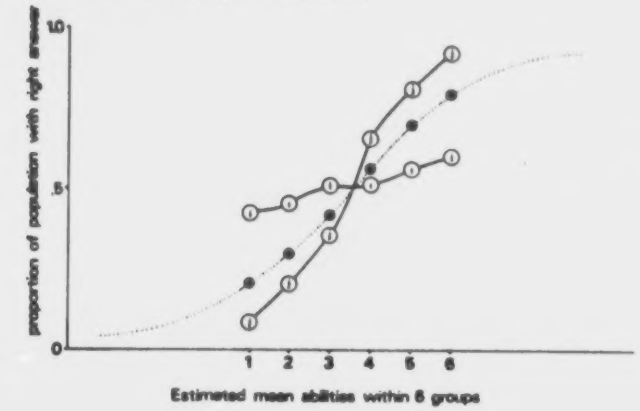
Figure 3: Item characteristic curve for item i



In fact this presents us with a rough estimation of part of the item characteristic curve (figure 3). Only a 'rough estimation', because no more than six points (i.e.: at the six ability levels) are defined, and only a 'part' because ability is not likely to vary from minus infinity to plus infinity among the testees.
This estimation of the probability of success is compared to the actually observed mean performance of each of the six groups on the item.
The difference between the estimated and the observed performance is the basis for the calculation of a measure of fit of the item in the test.
Of course there are numerous possibilities for departure of the observed performance from the estimated curve.

14

Figure 4: Departures from estimated curve for two items, i and j, of equal difficulty



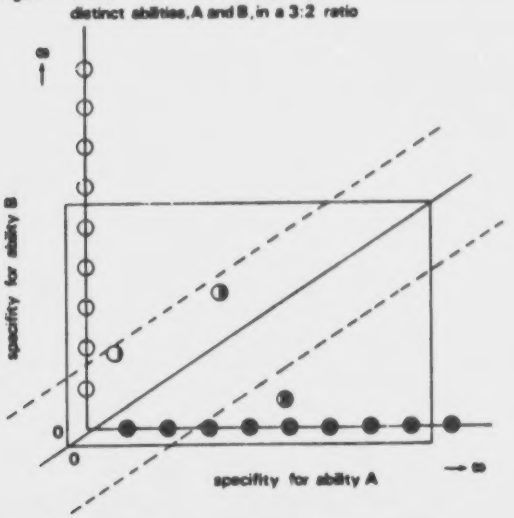
●·····● Estimated probability of right answers for items i and j at the ability level of 6 groups.
①—① Observed proportion of groups 1 to 6 with right answers on item i; weak discrimination between groups.
①—① Observed proportion of groups 1 to 6 with right answers on item j; strong discrimination between groups.

In figure 4 two of these possibilities are presented, for two items (i and j) of equal difficulty: the probability of success for each of the six groups is the same on both items. The items do not differ much in degree of departure from the expected curve thus their measure of fit in the test will not differ much. But on item i a larger number of persons from the low ability groups gave a right answer than was expected, whereas in the higher ability groups there are fewer right answers than expected. The item discriminates less well between the ability groups than we would expect from their performance on the whole test. For item j it is just the opposite: this item is highly discriminative.
In fact, the difference in ability between the six groups is invariable, or at least is not likely to change during the administration of the test, thus the differences in performance between the groups should remain constant on the items in the test if the items all measure the same ability.
Now in postulating unidimensionality of the ability, what we expect in general is that whenever an item does not fit, i.e. does not measure the same differences in ability between the groups of the target population as the whole test does, the item requires a different ability and natives will not necessarily show high scores on this item; they might even obtain lower scores than non-natives.

However, imagine a test constructor designing a test to measure physical strength. As items he uses different sizes of nails. He postulates: large long nails require more strength to hammer down than small, thin ones do. He scores the persons taking his test by the number of blows they need to drive in each nail completely. Obviously, if he presents his test to people who are poor at aiming, he

is liable to find physical strength to be higher correlated with number of smashed fingers than with number of blows. What he is doing is measuring two distinct traits or abilities at the same time: physical strength and skill at aiming.

Figure 5 : Latent trait of test T , requiring two distinct abilities, A and B, in a 3:2 ratio



● Items requiring ability A exclusively.
○ Items requiring ability B exclusively.
◑ Examples of items requiring both ability A and B.
▭ The test T .
/ The latent trait of test T .
--- Limits for items to fit in test T .

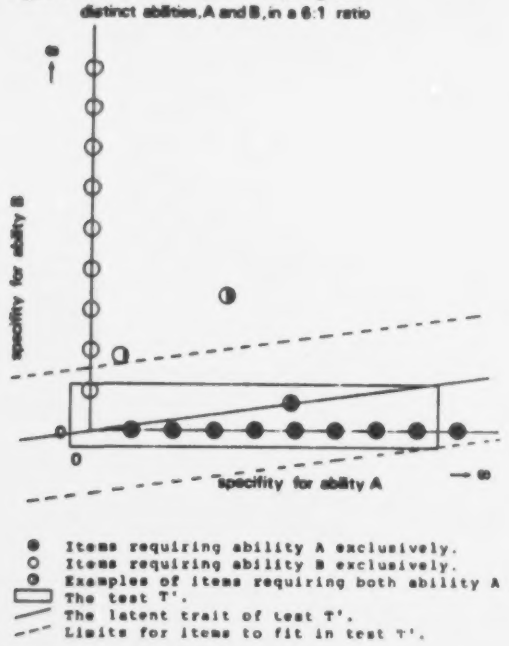
Figure 5 gives a graphic representation of such a test. Test T requires two distinct abilities A and B in a 3:2 ratio. The items requiring either ability A or ability B exclusively are represented as dots on two lines intersecting at an angle of 90°. Items at the intersection of the two lines require no specific ability, neither for ability A nor for ability B. Moving away from the intersection, items require more and more specificly either ability A or ability B and will consequently discriminate more on that ability between persons taking the test. Because the test requires both abilities A and B, the latent trait of test T will be defined as a line going through the intersection of trait A and trait B at an angle defined by the ratio of the two abilities in the test. For items to fit in the test they will have to test both abilities more or less in this ratio or one of either ability at a not too specific level. Items discriminating highly on either ability will not fit.

Because test T contains more items requiring ability A than items requiring ability B, highly discriminating items on the trait of test T are more likely to test ability A specificly, whereas items with low discriminative power in the test have a greater chance to be items testing ability B.

In the example of the test constructor who wanted to test physical strength the best solution for him is to get rid

of the tiny, thin nails that require a good aim.
What will happen with the test is shown in figure 6.

Figure 6: Latent trait of test T', requiring two distinct abilities, A and B, in a 6:1 ratio



● Items requiring ability A exclusively.
○ Items requiring ability B exclusively.
◑ Examples of items requiring both ability A and B.
▭ The test T'.
/ The latent trait of test T'.
--- Limits for items to fit in test T'.

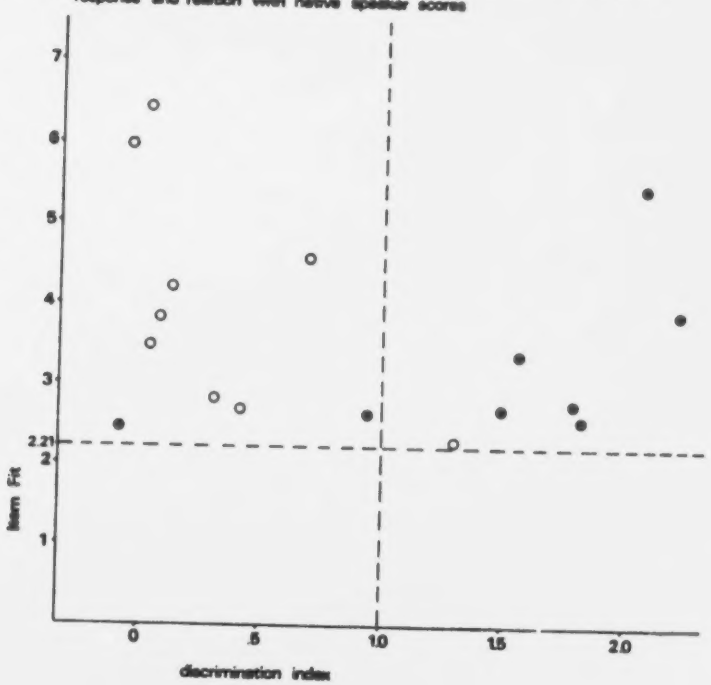
Test T' still requires the same two distinct abilities, but now the ratio is 6:1. A number of items requiring ability B have been deleted from test T and a new definition of the latent trait for test T' is obtained. The angle formed by the latent trait of test T' and ability A has become smaller and automatically more items requiring ability A will fit in test T'. Items with high discriminative power on ability A now fall within the limits for items to fit in test T'.

With this picture in mind we can redefine our expection mentioned earlier: whenever an item does not fit in our listening comprehension test and this item has a low discrimination index, we do not expect native speakers to do well on this item. In fact, they might just as well score lower than (some of) the non natives from the target population. If, however, an item does not fit, and misfit is due to high discriminative power on the trait of the test, we expect that this item is no problem at all for natives; the item will eventually fit after the items showing misfit in combination with a low discrimination index will have been deleted from the test.
In other words: we expect to find a correspondence between native speaker ability and the latent trait observed in Rasch analysis of foreign language learner response, and to be able to improve this correspondence by deleting items on which low native speaker scores were observed.

Results

In figure 7 the discrimination index of the misfitting items of the listening comprehension test is plotted against the measure of fit of the items. The dotted horizontal line represents the critical upper limit of the fit-measure (.05 level of significance in the F-distribution, with 5 over infinity degrees of freedom). The vertical line represents the median discrimination index: items to the right have more discriminative power than the test in general and items to the left are less discriminative.

Figure 7: Discrimination (x-axis) versus Fit (y-axis) in Rasch analysis of target population response and relation with native speaker scores



● prop. of natives with right answers > .80

○ prop. of natives with right answers ≤ .80

Each dot in the figure represents one of the misfitting items. In addition, for each item an indication is given whether more or less than .8 of the group of native speakers chose the right answer. We take .8 of the native speaker group as the lower limit for items that present no difficulties to the defined group of native speakers. The items tend to form a V-shape: the more the items deviate from a discrimination index of 1.0 the less well they fit in the test. From the picture it is clear that in general highly discriminative misfitting items gave less trouble to native speakers than low discriminating misfits.

Now we feel confident to start the operation of pulling out nails, i.e.: deleting items from the listening comprehension test.

Figure 8 A and 8B: Native speaker response versus measure of fit in Rasch analysis of target population response (59 items)

Figure 8A

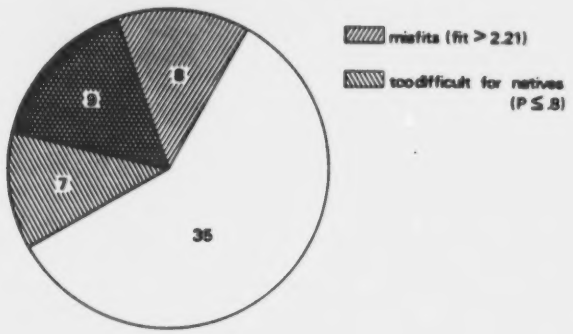
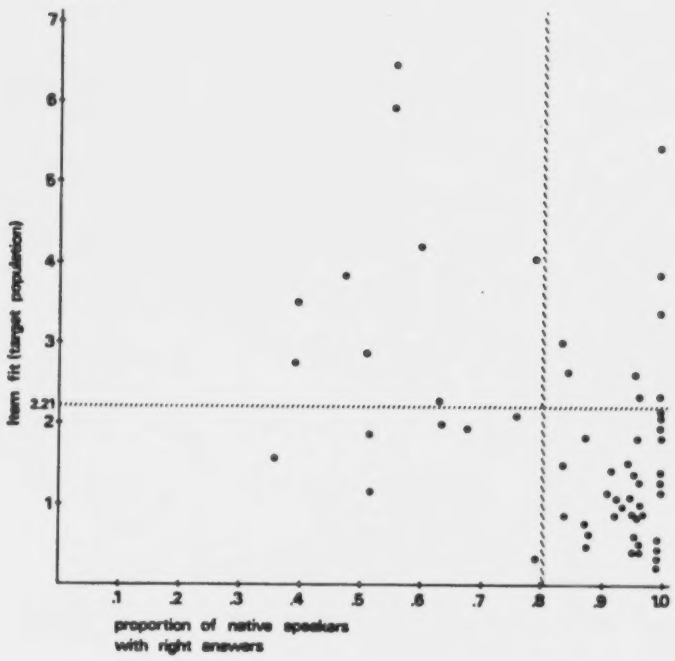


Figure 8B



In figure 8B every dot represents one of the 59 items in the listening comprehension test. The proportion of native speakers who gave right answers on each item (x-axis) is plotted against the measure of fit in Rasch analysis of the responses observed in the target population (y-axis). The critical upper limit of the fit measure at 2.21 is indicated by a horizontal line. The vertical line represents what we take as the lower limit for native speaker response. Most of the items (35 out of 59) fall within both limits. Another 9 items surpass both limits and thus confirm the hypothesis too: misfitting items are too difficult for natives because they measure the wrong trait, they do not test native speaker listening ability. A total of 44 (35+9) items come up to our expectation. Still we are left with 8 items showing misfit that are not too difficult for natives, and 7 items with a low native speaker score apparently fitting in the test: a total of 15 items contradicting the hypothesis. Figure 8A presents the same findings once

again: the complete circle represents the whole test of 59 items with the different sections representing the item falling either within or without one or both of our criteria.
Phi-correlation between measure of fit and native speaker score as dichotome variables is .37.

The next step is to delete the 9 items that surpass both limits and do another Rasch analysis on the remaining 50 items. The results are presented in figures 9A and 9B.

After the 9 items have been removed one more item falls within both limits and three more surpass both limits. We have gained four items: a total of 48 now comes up to our expectations. Phi-correlation has moved up to .56.

Figure 9A and 9B: Native speaker response versus measure of fit in Rasch analysis of target population response (50 items)

Figure 9A

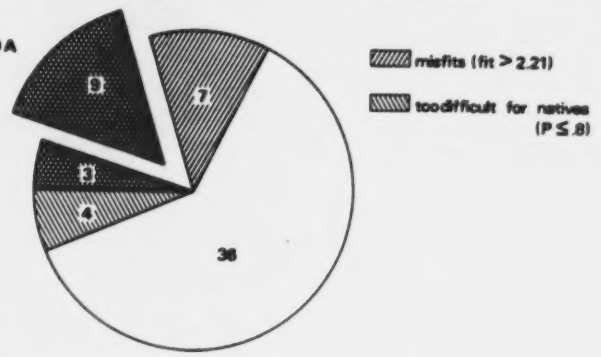
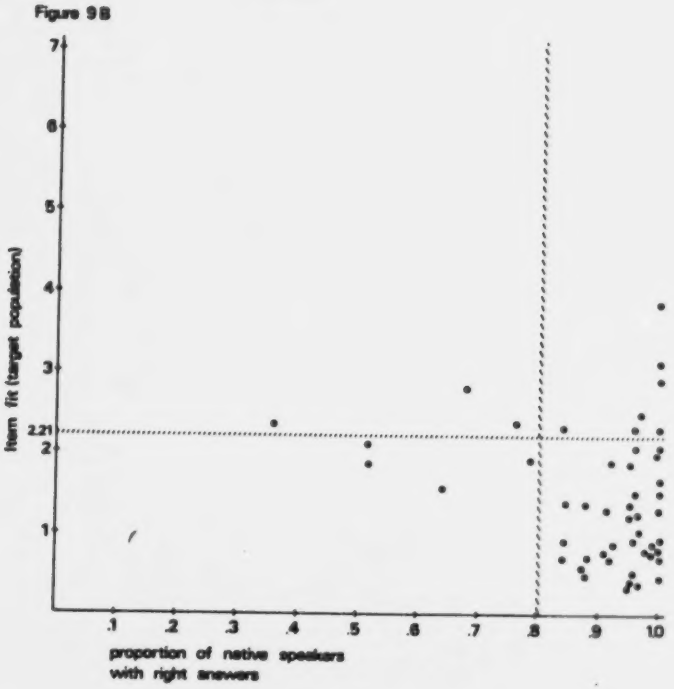


Figure 9B



Another 3 items can be deleted from the test now (figures 10A and 10B) leaving 47 items in the test. Two more items are now within both limits whereas 3 more surpass both limits: a gain of 5 items supporting the hypothesis. Only 6 items are left that do not correspond with our expectations. Phi-correlation is now .77.

Figure 10A and 10B: Native speaker response versus measure of fit in Rasch analysis of target population response (47 items)

Figure 10A

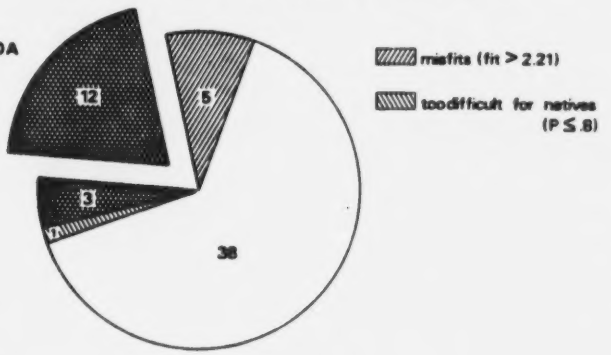
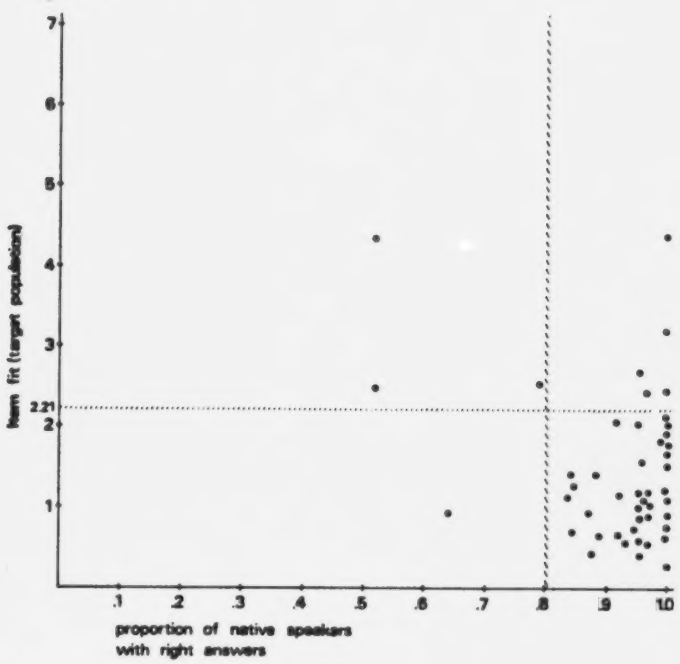


Figure 10B



Deleting the last three items that surpass both limits and reanalyzing the remaining 44 items (figures 11A and 11B) we find one more item to fall within both limits. There are no more items, however, that surpass both limits and the deletion process seems at a dead end. With a total of 54 items in favour of our hypothesis and 5 contradicting it, phi-correlation is now .80. But of the 4 remaining items that apparently misfit in spite of sufficient native speaker score, two items show a negative difference between the mean score of native speakers and that of the target population: the target population did better than native speakers on these two items. This means that the misfit of these items in fact does correspond with our expectation.

Figure 11A and 11B: Native speaker response versus measure of fit in Rasch analysis of target population response (44 items)

Figure 11A

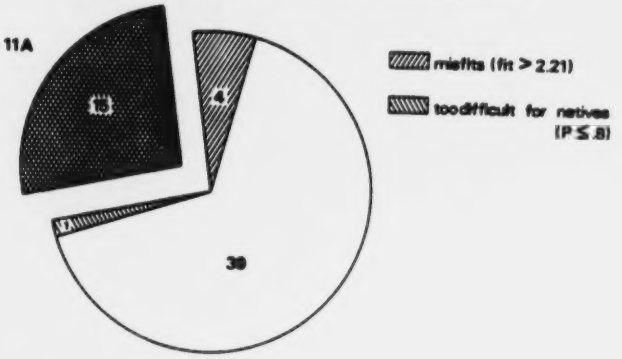
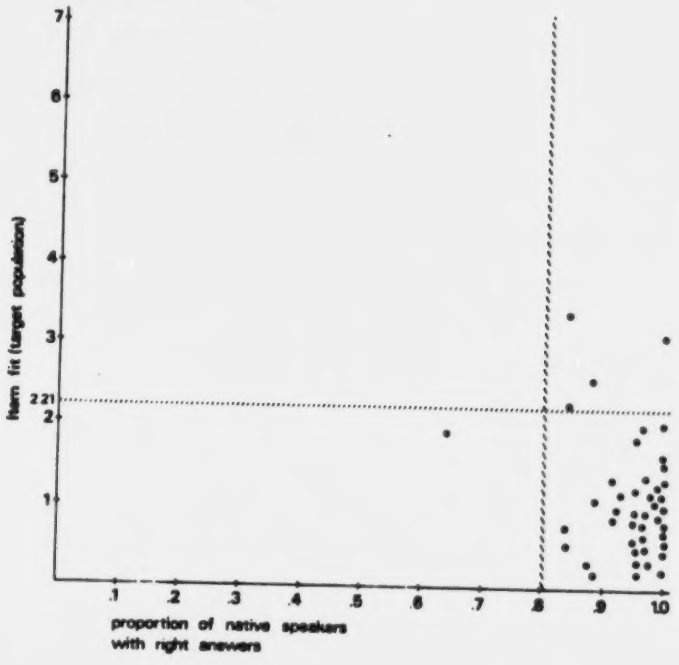


Figure 11B



Figures 12A and 12B show what happens if we delete these two items: one more item falls within both limits and only two out of the initial 59 items refuse to come up to our expectation: one slightly misfitting item easy enough for natives and one item that is too difficult for natives but still fits. The 40 items falling within both limits now cluster in the bottom righthand corner of the graph in figure 12B: these items fit extremely well in the test and present no difficulties for the group of native speakers.

Correspondence between native speaker ability and the latent trait observed in Rasch analysis of the target population response expressed in phi-correlation is .91, and confirms our hypothesis that the latent trait of this test is in fact the ability of native speakers to understand samples of their mother tongue. Thus 40 items out of the original 59 items will make a valid test for measuring listening comprehension of English.

Figure 12 A and 12 B : Native speaker response versus measure of fit in Rasch analysis of target population response (42 items)

Figure 12 A

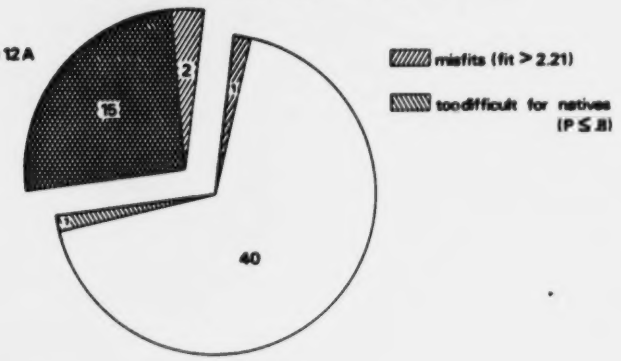
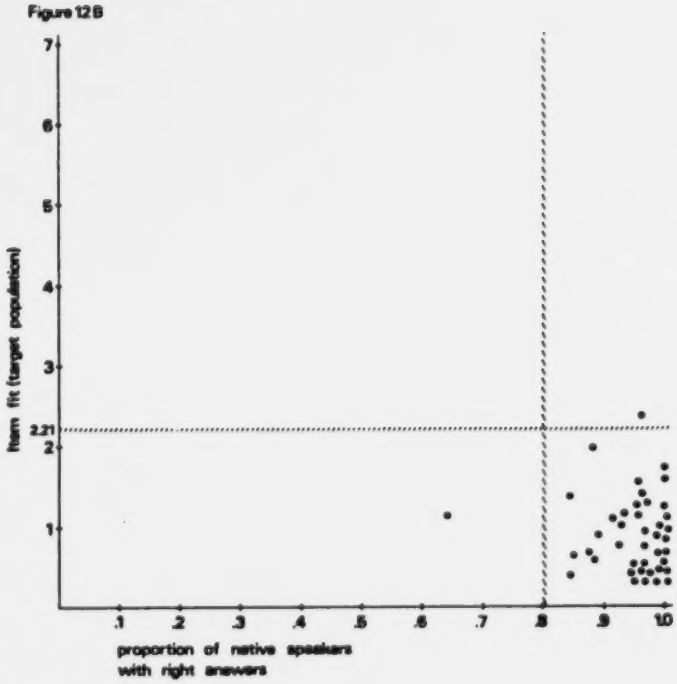


Figure 12 B



The remaining 19 items do not measure this ability. Figures 13A and 13B represent the results of a Rasch analysis of these 19 items regarded as a separate test. Figure 13A pictures the items in this test: 15 items surpassing both critical limits and 4 items surpassing only one of these limits: presented as a section from the orginial 59 item test (on the left) and regarded as a separate test (the circle on the right). Figure 13B shows that (apart from one slighty misfitting item) these items will fit together in a test. What this test measures is not clear as yet but anyhow it does not discriminate between able and less able listeners of English of comparable age and educational background.

Figure 13A and 13B: Native speaker response versus measure of fit in Rasch analysis of target population response (19 items)

Figure 13A

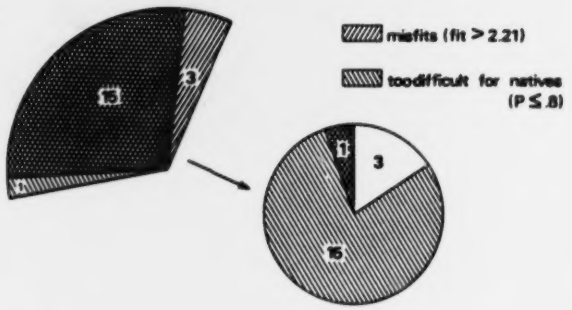
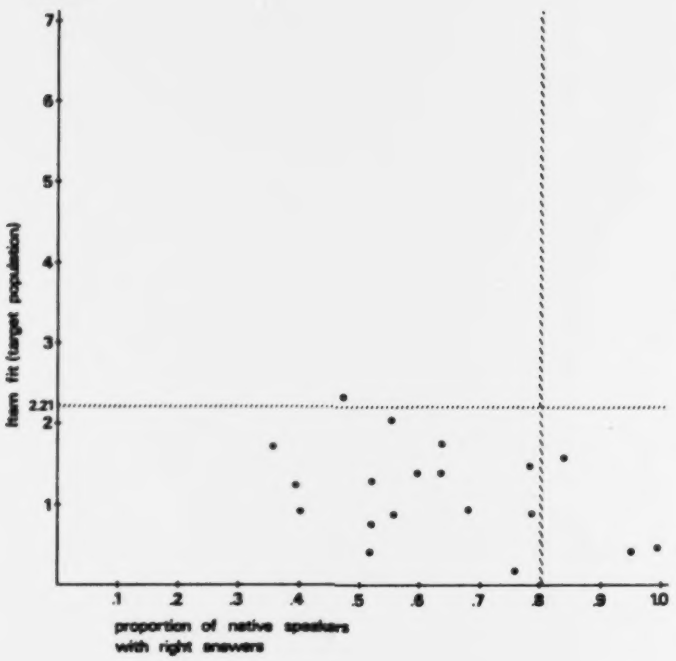


Figure 13B



Examples

I would now like to illustrate the different kinds of items in the test by giving examples of items that do fit in the test and of those that do not fit.

The first example is item number 17.

Item no 17

Tape:

(Wilson:) One of the things that worries me about them is how people are going to be able to keep track of their spending when they've just got a plastic card which they hand to the shop keeper. Now...

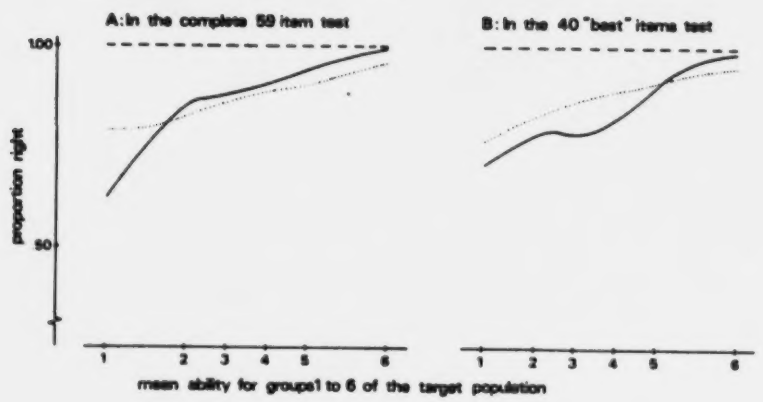
(Fortescue:) You've just raised a very important question, that if you're paying for goods with that wad of notes you can see how many you've spent and you have some idea of the total of your day's shopping whereas if you go into a number of eh shops and pay with eh the same credit card, you come away with nothing other than receipts and there's no sense of actually spending money.

Question booklet:

17 The use of a credit card can make people less aware of the amount of money they spend.
A True B False

The proportion of the native speaker group that chose the right answer (A) was 1.00 and .86 of the target population chose this answer. All the native speakers found the right answer obvious and for the target population it was an easy item. The item is easy <u>if</u> you understand English. Nevertheless, the item initially showed misfit (3.86) in Rasch analysis of the target population response. Misfit is combined with a high discrimination index of the item (2.35) but disappears after the operation of deleting items showing misfit along with a low native speaker score. Fit measure for item 17 improved and went down from an initial 3.86 to 2.97, 2.47, 2.10 respectively, as more and more items were deleted, in the way I have just described, ending up at 1.73 in the final 40 item 'good' test of figures 12A and 12B.
Figure 14 pictures the item characteristic curves of item 17 in the 59 item and in the 40 item test.

Figure 14: Item characteristic curves for item no 17


A: In the complete 59 item test
B: In the 40 "best" items test
100
50
proportion right
1
2
3
4
5
6
mean ability for groups1 to 6 of the target population
discr.: 2.36
fit: 3.86
discr.: 1.53
fit: 1.73
observed proportion of the native speaker group with right answers
expected proportion of 6 ability groups of the target population with right answers
observed proportion of 6 ability groups of the target population with right answers

Item no 14 is a different case:

Item no 14

Tape:

(Wilson:) Is it your impression though that eh people are in any way reluctant to use these new facilities?

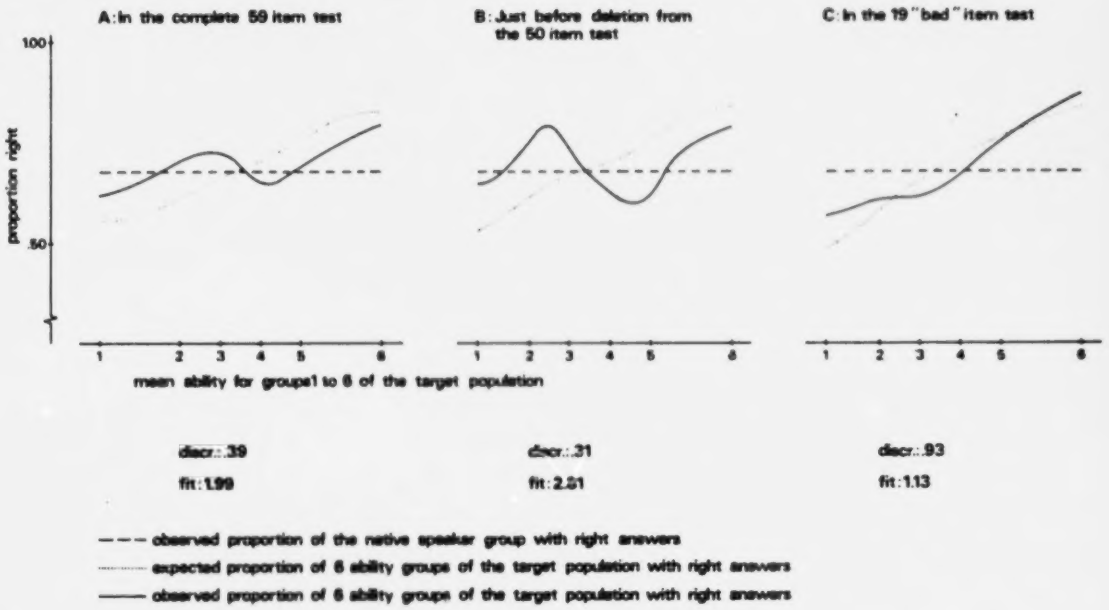
(Watts:) I think one of the problems is, that the credit card is an emotive term in some people's minds. An enormous number of people who have Barclay cards or Access cards, in fact use them just for the purchase of goods and there's no credit associated with it at all, and that's going to be true of the new payment systems.

Question booklet:

14 Some people do not like to buy things "on credit".
A True B False

The correct answer is A. A porportion of .68 of the native speaker group and .69 of the target population chose this answer.
Though .69 of the target population indicates that the item is of normal difficulty the native speaker score is far too low. The major part of the target population went 'through the ceiling' scoring higher than the native speakers. What is wrong? Mr. Wilson asks whether people are 'reluctant' to use the new facilities like credit cards etc. Watts's answer seems to be 'yes' because he gives an explanation saying 'credit is an emotive term'. Linking this to 'reluctance' used in the question, the listener can conclude that people do not like the idea of credit. The item is more a test of intellectual gymnastics than a straightforward test of English which explains why it is equally difficult for native speakers and for the target population who have a simular educational background. The item, however, fits in the 59 item test (1.99) but has a very low discrimination index (.39). After deleting the first 9 items in the operation described, fit measure leaps up to 2.82 and the item no longer fits.
Figure 15 gives the item characteristic curves of this item in the 59 item test (figure 15A), after deletion of 9 items (figure 15B) and in the 19 item bad test (figure 15C).

Figure 15: Item characteristic curves for item no 14



Both these examples were true-false items. Item 43 is a modified cloze item with two options:

Item no 43

Tape:

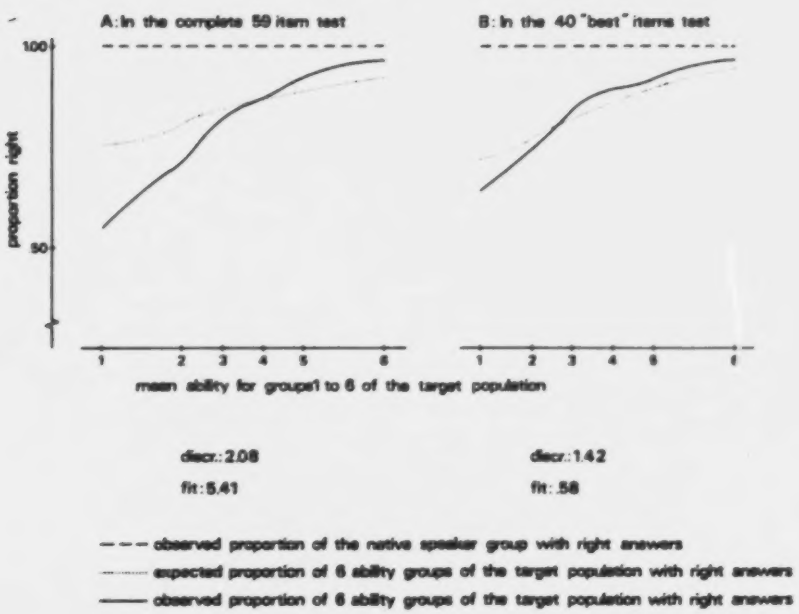
Prince Charles has been speaking about Britain's industrial future. He told delegates at a conference in Bournemouth of the iron and steel trades union that if only the two sides of industry would co-operate more, Britain's performance internationally could be ---Buzz---

Question booklet:

43 A vastly improved
B very disappointing

Item 43 is a good item: 1.00 of the native speakers and .86 of the target population chose the right answer (A). It is an easy item but not too easy. Nevertheless, the item does not fit (5.41) as discrimination is too high (2.08). In the subsequent stages of deletion the fit measure goes down from the initial 5.41 to 3.85, 3.21, 1.86 to a final .58 in the 40 item test: an extremely good fit! The discrimination index is now 1.42. Figure 16 shows the item characteristic curves in the 59 and in the 40 item test.

Figure 16: Item characteristic curves for item no 43


A: In the complete 59 item test
B: In the 40 "best" items test
100
50
proportion right
1
2
3
4
5
6
mean ability for groups1 to 6 of the target population
discr.: 2.08
fit: 5.41
discr.: 1.42
fit: .58
observed proportion of the native speaker group with right answers
expected proportion of 6 ability groups of the target population with right answers
observed proportion of 6 ability groups of the target population with right answers

We continue with item no 44:

Item no 44

Tape:

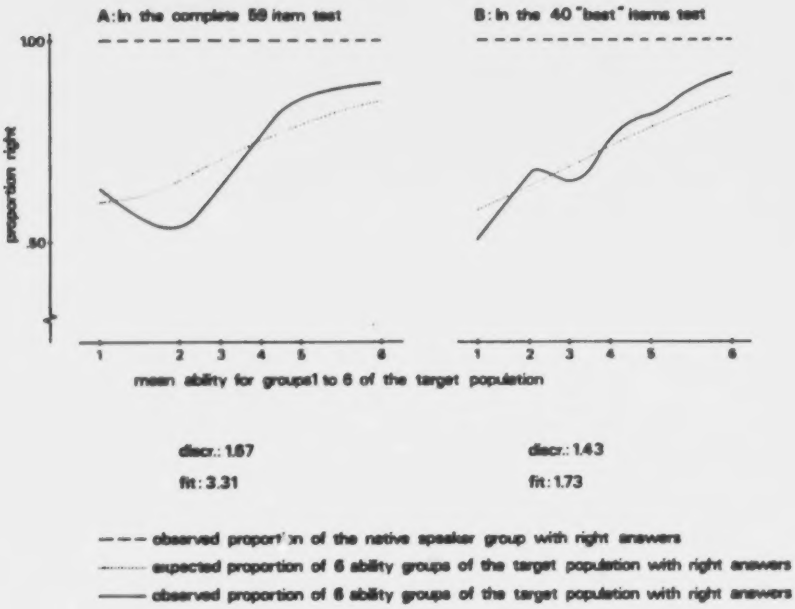
Prince Charles appealed for management and unions to reach a better understanding and went on to express his concern that Britain was ---Buzz--- its overseas competitors.

Question booklet:

44 A tolerating
B falling behind

Again this item was no problem for the native speakers: 1.00 of this group gave the right answer against .71 of the target population.
Initial misfit of the item in the 59 item test is 3.31 but goes down to 1.73 after the deletion operation.
Figure 17 gives the item characteristic curves.

Figure 17: Item characteristic curves for item no 44


A: In the complete 59 item test
B: In the 40 "best" items test
100
.50
proportion right
1
2
3
4
5
6
mean ability for groups1 to 6 of the target population
discr.: 1.87
fit: 3.31
discr.: 1.43
fit: 1.73
observed proportion of the native speaker group with right answers
expected proportion of 6 ability groups of the target population with right answers
observed proportion of 6 ability groups of the target population with right answers

Item no 6 is another example of a trickey true-false item:

Item no 6

Tape:

(Wilson:) Seymour Fortescue, is this evolution to the elimination of money, one which is happening virtually all over the world?

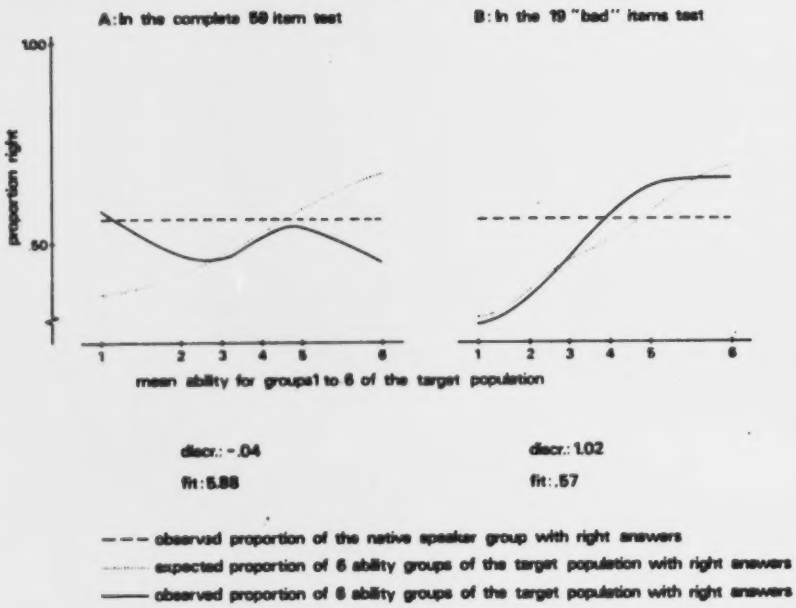
(Fortescue:) Yes I think it is. Perhaps the most interesting place is Japan, because the individual doesn't use cheques in Japan, they use cash and they use pass-books. You can see Japan making the switch from cash to electronic funds transfer without passing through the intermediate stage of cheques.

Question booklet:

6 Japan has already switched from cheques to electronic money transfer.
A True B False

Obviously, native speakers had great trouble in spotting the incorrectness of the statement: only .56 chose answer B. Of the target population a proportion of .52 chose this answer. The question is too difficult; the interpretation process necessary to arrive at the right answer is impossible in the short time allowed by the speed of the ongoing tape. The tone of voice is positive: there is 'elimination of money' in Japan, Japan is ahead in modernizing the payment systems but just because Japan is so far ahead and skipped the stage of cheques altogether the answer should be 'False'. The item shows severe misfit (5.88) and does not discriminate at all according to expectation among the testees in the target population (-.04)! In the 19 item 'bad' test, however, the item shows perfect fit (.57) and normal discrimination (1.02). It seems to me that the 19 item 'bad' test would be a good test of intelligence or alertness but it is not really a proper test of English listening comprehension.
Figure 18 shows the item characteristic curves of this item in the 59 item test and in the 19 item 'bad' test.

Figure 18: Item characteristic curves for item no 6


A: In the complete 59 item test
B: In the 19 "bad" items test
100
50
proportion right
1
2
3
4
5
6
mean ability for groups1 to 6 of the target population
discr.: -.04
fit: 5.88
discr.: 1.02
fit: .57
observed proportion of the native speaker group with right answers
expected proportion of 6 ability groups of the target population with right answers
observed proportion of 6 ability groups of the target population with right answers

The last example I wish to discuss is item 45, one of the two with a high native speaker score but an even higher score in the target population, the last two items to be deleted.

Item no 45

Tape:

Uganda's new president, Mr. Godfrey Benaisa, has appealed to the people of the country to co-operate in building a prosperous country that respects the rule of law, the protection of human rights and the establishment
---Buzz---

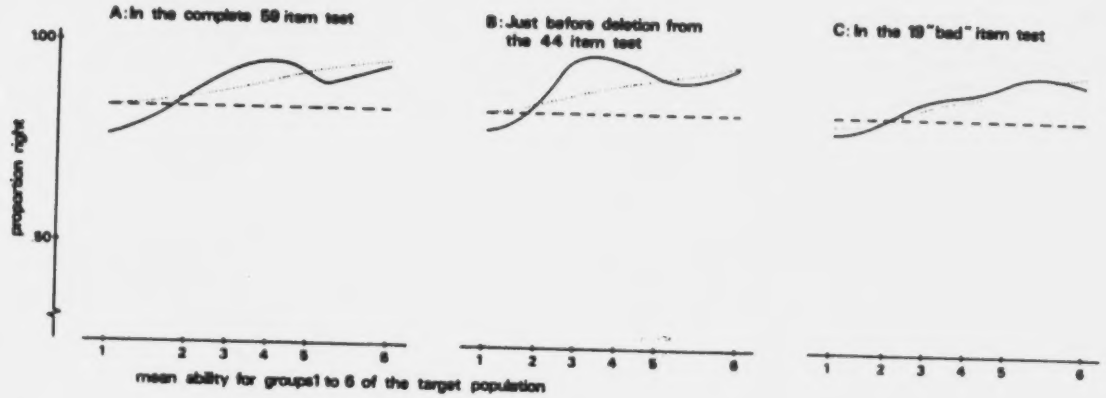
Question booklet:

45 A of a new state
B of democracy

Native speakers scored sufficiently high on this item: .84, but of the target population .90 chose answer B. One could defend both options, but the word 'democracy' seems to go better with 'respects the rule of law' and 'human rights' and was in fact the original word. Knowledge of political jargon in a revolutionary context and thus general intelligence or knowledge of the world is tested rather than English listening ability. The item fits in the original 59 item test but after several deletion operations it acquires a fit measure of 3.43 and is deleted in its turn. In the 19 item 'bad' test item 45

shows perfect fit.
Figure 19 gives the item characteristic curves.

Figure 19: Item characteristic curves for item no 45

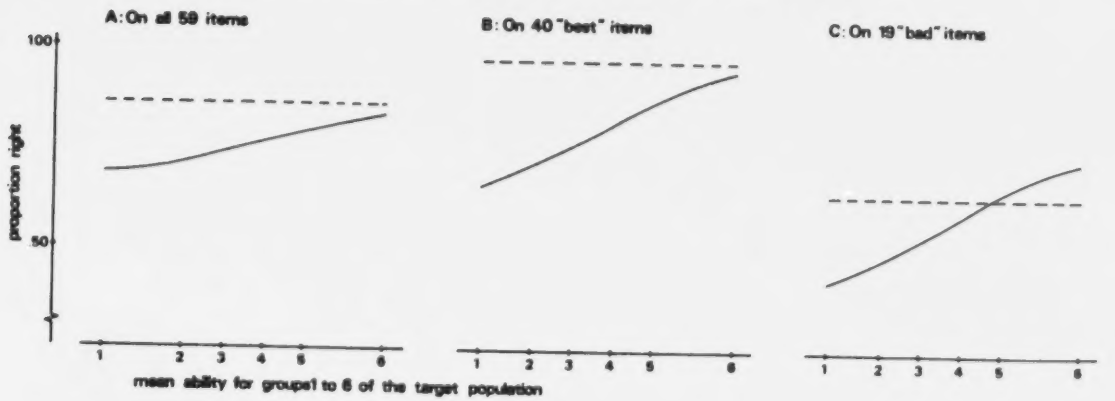

A: In the complete 59 item test
B: Just before deletion from the 44 item test
C: In the 19 "bad" item test
100
50
proportion right
1 2 3 4 5 6
mean ability for groups 1 to 6 of the target population

A	B	C
discr.: 1.29	discr.: .78	discr.: .95
fit: 1.48	fit: 2.83	fit: .15

---- observed proportion of the native speaker group with right answers
..... expected proportion of 6 ability groups of the target population with right answers
—— observed proportion of 6 ability groups of the target population with right answers

Figure 20 presents the mean of all item characteristic curves and shows the achievements of native speakers and of the six ability groups of the target population on the whole test and on selections of items from the test.

Figure 20: Mean scores in proportions of native speakers and of 6 ability groups of the target population


A: On all 59 items
B: On 40 "best" items
C: On 19 "bad" items
100
50
proportion right
1 2 3 4 5 6
mean ability for groups 1 to 6 of the target population

	A	B	C
native speakers	.86	.96	.63
target population	.73	.78	.58
difference	.12	.18	.05

---- observed proportion of the native speaker group with right answers
—— observed proportion of 6 ability groups of the target population with right answers

From this picture it is clear that the native speaker group obtained a higher score than the average testee from the target population, higher even than the average testee in the most able group (6) of the target population. The selection entitled '40 best items' creates a larger distance between the native speakers and the average of the target population. Moreover the 40 item selection discriminates better between the different ability groups in the target population. In the 19 'bad' items test discrimination between the ability groups of the target population is about the same but native speakers score only slighty above the average of the target population, native speakers have no advantage in doing this test: it is not a test of English listening comprehension for this population.

One point remains to be made clear. The six ability groups in the subsequent Rasch analyses are redefined for every seperate analysis: the testees are divided into six groups according to their score on the test that is being analyzed. The most able group on the 19 item 'bad' test is different altogether from the most able group on the 40 item 'best' test. To illustrate this, product-moment correlations were calculated between the 19 item 'bad' test and two seperate random selections without replacement of 19 items from the 40 item 'best' test. Selections were made to be of equal length for easy comparison.

Table of product-moment correlations between 3 selections of 19 items from the listening comprehension test (n=575).

	All	A	B	C
All 59 items	x			
A: 19 'bad' items	.40	x		
B: 19 'best' items	.78	.06	x	
C: 19 'best' items	.75	-.01	.47	x

From the correlation indices in the table it is clear that there is no relation between the 19 'bad' items and either of the two selections of 'best' items and testees scoring high on the 'best' items will not necessarily score high on the 'bad' items. The two selections of 'best' items have a positive correlation significant at the .005 level which is fairly high if we consider test length.

Conclusion

The latent trait of the listening comprehension test analyzed here can be identified as the ability native speakers demonstrate in understanding spoken samples from their mother tongue. Two thirds of the test constitute a valid measure for listening comprehension of English as a foreign language. The other third of the test seems more suitable for testing an ability which might be general intelligence or knowledge of the world but does not discriminate between different ability

levels of English listening comprehension.
A more general conclusion could be that Rasch analysis can be of great help in selecting valid items from a try-out version of a test. Provided that the major part of the try-out version does test the ability aimed at, the test constructor, by deleting misfitting items with a low discrimination index in successive Rasch analyses, can make a selection of items that have the highest probability of testing the ability he is aiming at.

American Psychological Association (1974) *Standards for educational and psychological tests*. F.B. Davis (chair), American Psychological Association, Washington, D.C.

Cronbach, L.J. and P.E. Meehl (1955) Construct validity in psychological tests, *Psychological Bulletin* 52, 281-302. Reproduced in Noll, V.H., D.P. Scannel and R.P. Noll (eds) *Introductory readings in educational measurement*, Houghton Mifflin, Boston 1972, 90-121.

Jong, J.H.A.L. de en H.W.M. van den Nieuwenhof (1982) *Een experimentele luistervaardigheidstoets*. Specialistisch Bulletin no 14, Cito, Arnhem.

Vollmer, H.J. (1981) Receptive versus productive competence? Models, findings and psycholinguistic consideration in L2-testing. Paper presented at the Sixth International Congress of Applied Linguistics, AILA 81, Lund 1981.

Vollmer, H.J. and F. Sang (forthcoming) Competing hypotheses about second language ability: a plea for caution. To appear in J.W. Oller (ed) *Issues in language testing research*, Newbury House, 9 Rowley, Mass.

Wright, B.D. and R.J. Mead (1975) *Calfit*: Research Memorandum number 18, Department of Education, University of Chicago, Chicago.

Wright, B.D. and M.H. Stone (1979) *Best test design: Rasch measurement*, Mesa Press, Chicago.

The traditional multiple choice reading comprehension test, used in the Dutch final examination, has always met with all sorts of criticisms. One of them is, "Our pupils have not really understood what the passages in the test are about, even though their scores on the multiple choice questions are satisfactory!"
We have tried to find out if there really is a different skill involved when pupils answer multiple choice questions in a reading comprehension test compared to when they summarize the reading passages used in the same test.

When the girl who was going to type the abstract for this paper asked me what title I wanted to give my paper, I perhaps too rashly thought of 'Rasch and Reading'. I am of course not going to deal with reading as a whole. Nor with reading in the mother tongue. What I do want to tell you a little about is about reading in a foreign language, or rather, about three ways of testing reading comprehension in a foreign language, and the analysis of the results of these tests. First, some background information. Here, at CITO, a lot of time and money is spent on the construction of multiple choice foreign language reading tests. They play an important role in the final examinations of secondary schools. Illustrative is the fact that a pupil's result on the test is weighted 50% in his examination mark. So there is plenty of reason to be very alert to criticism, and to try and improve the test. Over the years we have had quite a lot of it: most of it having to do with particular items or tests, or question types, but a considerable part of it was criticism on the method we used. So one important area needing attention is that of content validity. We have to see to it that the test measures the skill it is intended to measure. This is more easily said than done, of course. We do not have explicitly defined objectives for the required levels of reading comprehension. We only know that we should provide 5 or 6 reading passages, followed by 50 multiple choice items testing whether the candidate has understood the passage. There is no vocabulary-list we have to stick to. We do not work according to a matrix or hierarchy, but formulate questions always bearing in mind: is this the information the writer of the article would also call relevant? Is this what he wanted to point out to us? This method, though not very scientific or exact, produces question types that could be fit into a matrix, in other words: the questions do seem to appeal to various subskills of the reading skill. So there are questions asking the candidates to reproduce, draw a conclusion, recognize the tone or style, state the main point. All in MC, of course, so the candidate has to recognize a good reproduction, conclusion, etc. The question is how to prove that the test is valid. We have tried to do some research in various ways.

First we tried factor analysis. We factor-analyzed the results on subskill-tests and the traditional multiple choice test. This confirmed the hypothesis that the traditional test appeals to a considerable number of subskills of reading comprehension.

In the second stage we compared the traditional reading comprehension test to two other forms of multiple choice reading comprehension test, this time by means of Rasch analysis. I shall not go into the results of the second stage. It made us only conclude that the trait or skill - the reading skill - did not differ in the three tests. We could not say that any test was better than or preferable to the other two tests. The analysed data in the third stage however, gave us a little more food for thought.

In the third stage we decided we would like to be able to confirm or deny the repeated criticism of teachers that their pupils cannot really tell what the passages in the test are mainly about, even though their scores on the multiple choice test are satisfactory. In other words, candidates do know how to choose the right answer in multiple choice items, but they would not be sufficiently able to formulate in their own words the right answer to an open-ended question, and, above all, they would not be able to summarize the passages in the test.

So what we did was: we chose two passages from a traditional multiple choice examination test of English reading comprehension. Together with the original items this would be test number one. Test number two consisted of the same two passages, this time followed by open-ended questions. The questions were more or less the same as the stems of the multiple choice items in test 1. Test three consisted of again the same two passages, which the candidates had to summarize in their own words, and in Dutch. There was a maximum number of words. We were of course aware that tests 2 and 3, with their 'open' nature would cause problems. They would have to be rated in such a way that each candidate would get a fair treatment. We spent a lot of energy on the response models, the way of scoring the answers, pretesting etc.

I have now come to the method we used to analyse the data we obtained after the results of the tests had been returned to us. You may already have guessed: Rasch's latent-trait analysis. It permitted us to join the three tests in one model. This would enable us to see whether the tests fitted the model, in other words, we could find out whether the tests measured the same skill. Moreover, the difficulty parameter would show whether the three types of tests differed in difficulty. So with this method we could ask ourselves the questions:

- do the tests fit the model? or perhaps which test fits the model best?
- what are the differences in the difficulty parameters?

Now I would like you to have a look at the handouts. If you look up page 43 in the appendix, you will see the fit analysis. The three tests are considered to be one test of 44 items: the first column on page number one shows the range of scores, split up in 6 scoregroups. Group one has

the lowest scores, group 6 the highest. So the first column deals with item difficulty; it gives it the item characteristic curve. The second column, again consisting of the 6 groups, shows us the number of unexpected answers. For every item and every group the difference between what has been observed and what should have been observed is calculated. The difference is made visible by means of stating the number of too many or too few right answers.

The third column gives the model fit, per group of scores and per item.
The last column shows the model fit per item, after all the groups have been taken together. It shows what is called the fit mean square. Looking at the fit mean square is the quickest way to find out which items do not fit into the model, or to be more positive, to see if the test is homogeneous. Before telling you our conclusions from these data I should point out four causes that could lead to lack of fit. First, **heterogenity of items.** Speaking in terms of factorial itemstructure this would mean, that more than one factor, one variable, is necessary to describe the test items.

Secondly, lack of so called local stochastic independence could lead to lack of fit. This may occur when a variable, not permitted by the Rasch model plays a certain role. This happens for instance when a candidate's memory helps him to find the right answer.

A third cause may lie in persons guessing the right answer. The Rasch model does not allow guessing, but in reality, in a testing situation, it is very difficult to eradicate guessing. When guessing has been frequent, one will find in the lowest scoregroups an overrepresentation of right answers with the difficult items and vice versa an underrepresentation with the easy items. Of course in the case of open-ended questions and a summary guessing cannot play a role.

A fourth cause could be 'item bias'. This occurs when certain items have systematically been too easy or too difficult for one or more identifyable subgroup within the sample.

In order to explain lack of fit, one should first think of lack of local stochastic independence, guessing, and item bias. If these cannot explain the items in question, then item-heterogenity is to blame.

Now, do we in fact have to think of causes of lack of fit in this particular case? Are there many items with a bad fit? Not too many. I have indicated on the analysis which items formally do not fit: numbers 6, 9, 14 and 43.

So our most important general conclusion from this Rasch analysis is that there is one trait, measured by all three different reading comprehension tests. The three tests deal with the same skill. This is of course an agreeable conclusion to come to, if one, for various reasons, would like to stick to the traditional multiple choice test in the final examinations. But if one looks

at the data in detail, there is some more to be said. I'll come back to this later. First I would like to discuss shortly the 4 items that had bad fit.

What could have caused them to be ill-fitting? It is remarkable by the way, that 3 of these 4 items are in the multiple choice test.

A close look at these 3 items shows that there are flaws within the items. They are either too confusing, adding the variable: which I call puzzle(-mindedness), or they require the Dutch pupil to understand typically English ways of expressing feeling or attitude such as when an Englishman uses an understatement. So in the first case we think the bad fit may have been caused by guessing: the candidates just did not know what to choose. The lack of fit in the item which required an understanding of the English way of expressing feeling or attitude may be due to heterogenity as there may be an additional trait involved: a trait having to do with the personality of the candidate being tested.

The fourth item, number 43, in the summary, may simply have been too difficult, requiring too much knowledge of English of the candidates concerned. It is interesting to note that knowledge of English seems to become a separate variable, once a certain level of difficulty has been passed. Before passing that level it might be reading comprehension that matters, and not reading in a foreign language that matters.

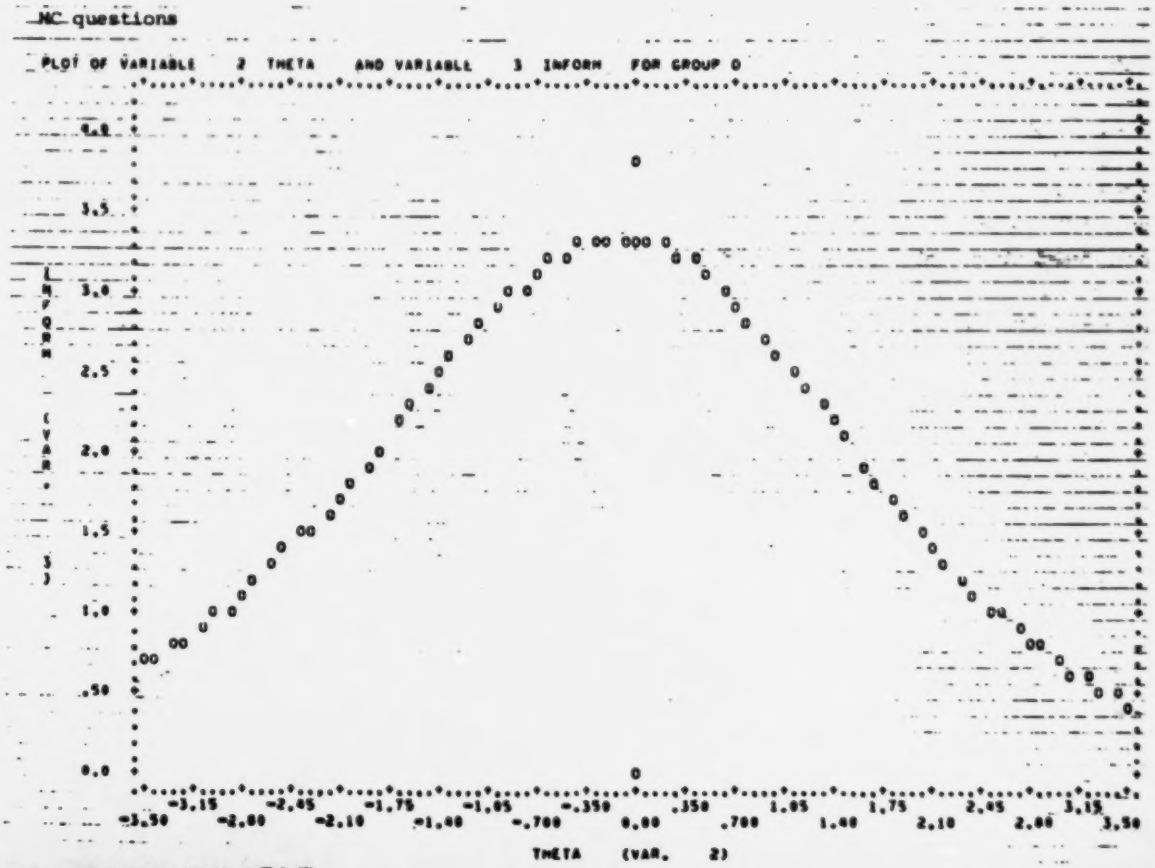
All in all, there are only four items deviating from the test do fit the model. Our first question was answered. You may remember that we also wanted to have a look at the difficulty parameters. They show that the tests do differ in difficulty. The multiple choice test has been the easy one, followed subsequently by the open-ended questions and the summary. In other words: the more open the test is, the more difficult it is. So in a way the criticism, that pupils can answer multiple choice questions correctly, but are not able to formulate answers or give a summary, is understandable. But it is just a question of difference of difficulty of the tests, not of skill of the candidates.

Is its level of difficulty one more reason to stick to the traditional multiple choice test? I suppose it is not, or should not be. Moreover, I have not yet shown you the information curves of the three tests. We need to look at them too in order to find out the accuracy with which the test measures. You can find the curves on page 44. The ideal situation would be if I could also show to you the so called true ability distribution but certain circumstances have made this impossible. The horizontal axis shows ability, the vertical axis shows information. The majority of the candidates is to be found in the area from .350 in each graphic. Comparison of the three curves shows that compared to the i.c. of the MC test the information curve of the summary descends more slowly. This means that of the two the summary measures most accurately, and especially where candidates with the highest ability are concerned. This might be one point in

favour of the summary, if the purpose of the test is maximally accurate measurement in the area where most candidates are to be found. So here the analysis gives us a hint as to what test to prefer, or what direction out research should take. The open-ended questiontest does not do too badly either, of course.

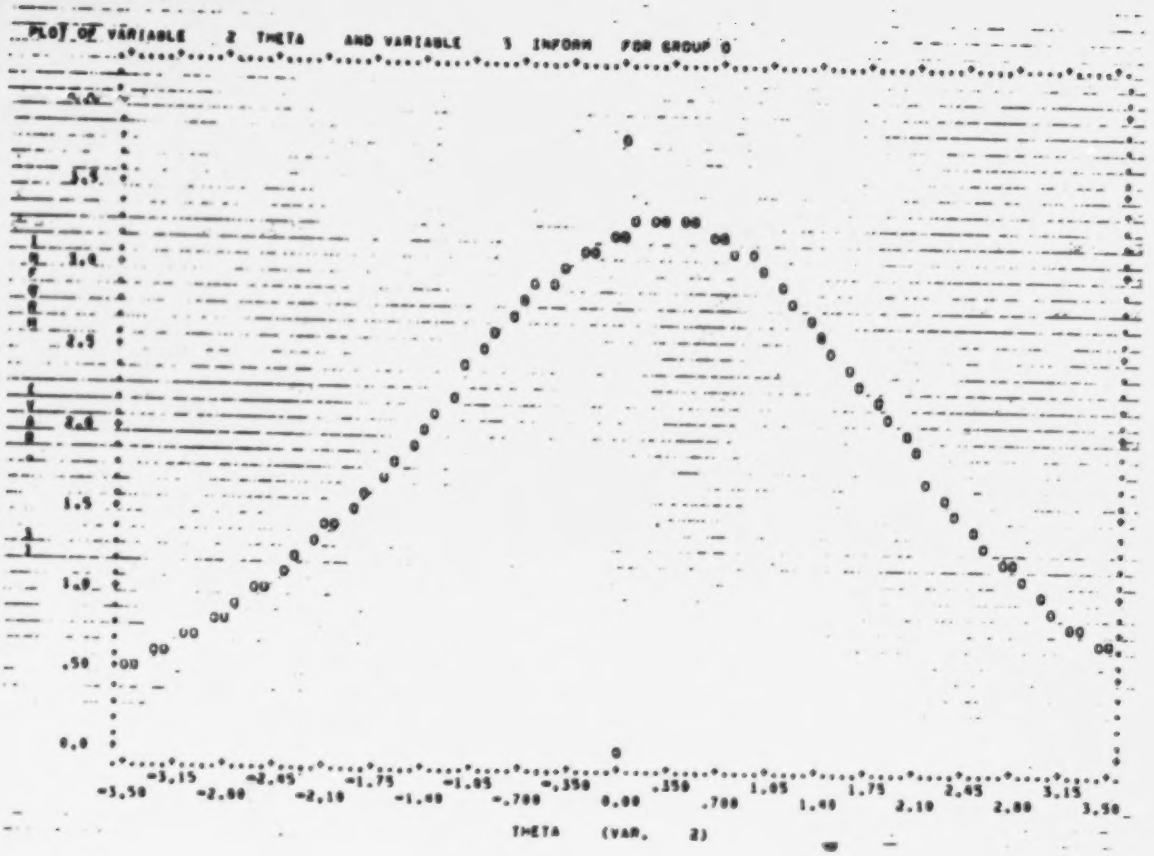
Anyway, we have, for the next step in our research, decided to tackle the summary again. We do know that in the situation of a final examination there are various practical reasons not to choose a summary ever. The construction of the response model and rating, finding teachers who are willing to rate, are all hazardous and time-consuming activities. Even so, we want to have more information on the subject of the summary, as compared to the traditional multiple choice test. This time the passage to be summarized will be another one than in the traditional test. It will be longer, as we feel that this brings the candidate closer to the real life activity. Perhaps it will change the model fit. The passage will be tested in an English and a Dutch version, to find out what part knowledge of English plays.

	SEQ NUM	ITEM NAME	L	ITEM CHARACTERISTIC CURVE. 1ST GROUP	2ND GROUP	3RD GROUP	4TH GROUP	5TH GROUP	6TH GROUP	NUMBER UNEXPECTED ANSWERS 1ST GROUP	2ND GROUP	3RD GROUP	4TH GROUP	5TH GROUP	6TH GROUP	FIT CHI-SQUARED 1ST GRP	2ND GRP	3RD GRP	4TH GRP	5TH GRP	6TH GRP	FIT L MN.SQ
MC questions	1		1	.81	.97	.92	1.00	.97	1.00	-2.0	1.7	-1.1	1.4	-.1	.4	1.3	1.3	.8	1.8	.8	.7	1.03
	2		1	.88	.99	.89	.83	.76	.81	-3.7	1.9	-1.8	3.3	.2	-2.5	2.2	.6	.1	4.5	.6	1.2	1.08
	3		1	.90	.94	.82	.67	.85	.93	2.0	-1.9	-1.5	-2.0	2.1	1.7	.6	.5	.3	.7	1.0	.9	1.06
	4		1	.90	.96	.88	.78	.91	.96	-.1	-3.1	-1.6	.2	2.7	2.7	.8	1.5	.4	.9	1.9	2.0	1.12
	5		1	.72	.91	.92	.89	.82	.86	-1.2	-1.8	1.1	-1.9	1.4	1.3	1.8	.4	1.3	2.2	3.1	.4	1.51
	6		1	.87	.94	.88	.89	.87	.87	1.1	-.4	1.5	-1.5	-.1	-1.4	4.6	.4	4.8	1.8	.8	8.7	1.18
	7	L	1	.72	.82	.89	.94	.82	.98	.6	.1	1.1	1.4	-3.6	.4	1.1	.4	.8	1.3	5.4	.1	1.27
	8		1	.82	.86	.85	.93	.85	.90	2.1	.8	2.1	-.4	.8	-5.9	1.0	.1	.7	.1	.8	4.8	1.25
	9	L	1	.84	.89	.87	.87	.78	.76	7.5	4.8	-2.9	-1.5	-4.5	-4.8	8.8	3.8	1.1	.8	1.3	7.1	1.47
	10		1	.38	.89	.81	.78	.76	.83	-1.7	-3.8	5.9	2.3	-.4	-2.1	.4	1.5	9.1	.9	.1	2.3	1.55
	11		1	.84	.72	.86	.86	.88	.95	.1	-1.7	1.8	-.2	-.4	.2	.1	.5	.7	.8	.1	.8	.28
	12		1	.88	.86	.81	.87	.84	.98	-1.8	-1.4	-1.7	1.8	-2.2	5.8	.6	.3	.9	.3	.8	3.7	1.69
	13		1	.83	.82	.78	.87	.79	.84	2.6	.7	1.1	-2.4	.2	-1.5	1.0	.1	.2	1.0	.0	.2	.43
	14		1	.87	.89	.87	.89	.95	.98	2.2	2.4	-2.2	.1	-7.8	1.4	.7	4.1	.7	.0	9.7	.0	2.03
	15		1	.87	.89	.78	.81	.76	.79	1.1	-2.4	4.7	3.1	-.8	-4.2	.2	1.8	3.2	1.0	.1	5.3	2.04
open-ended questions	16		1	.91	.95	.92	1.00	.91	.98	1.0	1.0	5.9	1.3	-12.0	3.0	.8	.4	.3	20.7	4.1	.1	1.82
	17		1	.19	.21	.37	.44	.61	.71	-1.0	-4.9	6.8	-1.2	1.5	-1.1	.2	3.1	4.2	.2	.1	.2	1.56
	18	L	1	.38	.34	.44	.44	.64	.79	2.7	-2.5	-1.6	2.6	-.5	-.8	1.4	.8	.3	.8	.8	.1	.57
	19		1	.84	.44	.59	.64	.82	.95	1.3	-4.1	-1.0	-1.7	2.2	3.5	.1	2.2	.1	.1	.1	3.1	1.22
	20		1	.16	.15	.29	.31	.67	.37	.6	-1.6	-1.6	-1.8	7.7	-2.7	.0	1.6	.3	.5	8.0	1.0	2.07
	21	L	1	.47	.67	.68	.75	.91	.83	-.3	1.6	-.8	-.2	3.2	-2.9	.0	.1	.0	.1	.0	2.5	1.99
	22		1	.33	.62	.76	.86	.82	.95	-.1	-2.5	.6	2.4	-.9	1.3	.0	.9	.9	1.2	.2	.7	.58
	23		1	.83	.28	.28	.33	.38	.71	-3.3	2.7	-1.1	-.8	-3.6	6.1	3.5	1.3	.2	.0	1.5	2.3	1.55
	24		1	.38	.76	.82	.72	.73	.98	-2.2	5.9	-1.7	-.2	-4.0	.6	.0	4.6	.4	.0	.9	.1	1.15
	25		1	.31	.46	.81	.61	.76	.81	-2.0	-2.3	7.6	-2.0	.7	-2.1	.7	.7	8.4	.0	.1	1.1	1.93
	26		1	.90	.82	.82	.72	.82	.86	1.8	-.9	-1.7	-.2	1.0	-1.4	.3	.1	.4	.0	.2	.0	.31
	27		1	.87	.59	.82	.72	.79	.81	1.7	1.0	-.7	.6	.6	-2.9	.1	.1	.1	.1	.1	2.6	.53
	28		1	.89	.28	.36	.39	.65	.79	-2.5	.8	-1.5	-.6	-1.1	4.3	1.7	.0	.3	.1	.2	2.0	.84
	29		1	.38	.54	.59	.64	.76	.88	-.2	.5	-.8	-1.2	.6	.8	.0	.0	.0	.2	.1	.2	.99
	30		1	.38	.46	.39	.33	.67	.79	3.2	-.5	1.3	-1.4	-1.0	-2.4	.4	.1	.1	1.7	.2	.1	1.48
summary	31		1	.88	.59	.82	.81	.82	.98	-.2	.1	-1.7	2.8	1.0	3.6	.2	.1	.1	1.3	.2	.1	1.71
	32		1	.83	.28	.32	.53	.61	.76	-3.3	-.9	-1.9	2.9	4.5	1.9	6.0	.1	.5	1.2	.9	.0	1.60
	33		1	.31	.49	.69	.75	.76	.95	-2.0	-1.3	-4.2	3.0	.7	3.9	.7	.2	2.5	1.0	.1	3.7	1.88
	34		1	.63	.85	.78	.69	.79	.98	2.6	6.3	-1.6	-3.8	-2.0	-.8	1.1	5.9	.5	3.1	1.1	.1	2.38
	35		1	.84	.54	.82	.72	.73	.93	.7	-1.8	-.7	.6	-1.6	2.1	.1	.1	.1	.1	.0	1.1	.36
	36		1	.89	.77	.89	.86	.91	.86	1.5	.8	2.6	.1	.5	-3.8	.4	.0	1.5	.0	.1	9.0	1.60
	37		1	.86	.82	.81	.94	.82	.98	-.3	1.3	-.9	2.6	-2.9	1.0	.1	.3	.2	2.1	3.5	.0	1.12
	38		1	.38	.59	.57	.67	.73	.86	-.2	2.5	-1.6	-.2	-.9	-.2	.0	.0	.3	.0	.0	.1	.19
	39		1	.58	.79	.73	.81	.76	.98	-.9	4.8	-1.6	.5	-2.7	-.6	.1	3.3	.0	.1	8.1	.2	.98
	40		1	.89	.28	.38	.44	.48	.57	-2.0	1.6	-.3	2.5	1.0	-3.5	1.1	.3	.0	.9	.1	1.7	.71
	41		1	.25	.13	.32	.47	.48	.81	2.6	-6.6	-1.6	1.5	-1.0	4.6	1.0	6.3	.3	.3	.1	3.8	1.83
	42		1	.83	.88	.85	.86	.86	.26	.3	-1.5	.1	-.4	-1.0	1.4	.2	1.9	.0	.1	.4	.4	.69
	43		1	.88	.18	.16	.15	.45	.84	-3.2	.7	-2.8	-.5	6.9	4.6	6.2	.1	1.6	3.5	2.7	8.5	2.66
	44		1	.86	.26	.24	.22	.48	.62	-2.1	1.9	-.8	-3.7	2.8	.6	1.5	.7	.1	2.1	1.2	.6	.94
	SCORE RANGE			1=20	21=24	25=26	27=29	30=32	33=43	NO	32	19	37	36	33	42							MEAN
	MEAN ABILITY			-.55	.66	.38	.67	1.09	1.87		PLUS=TOO MANY RIGHT MINUS=TOO MANY WRONG						1.9	1.6	1.7	1.0	2.2	2.0	0.0

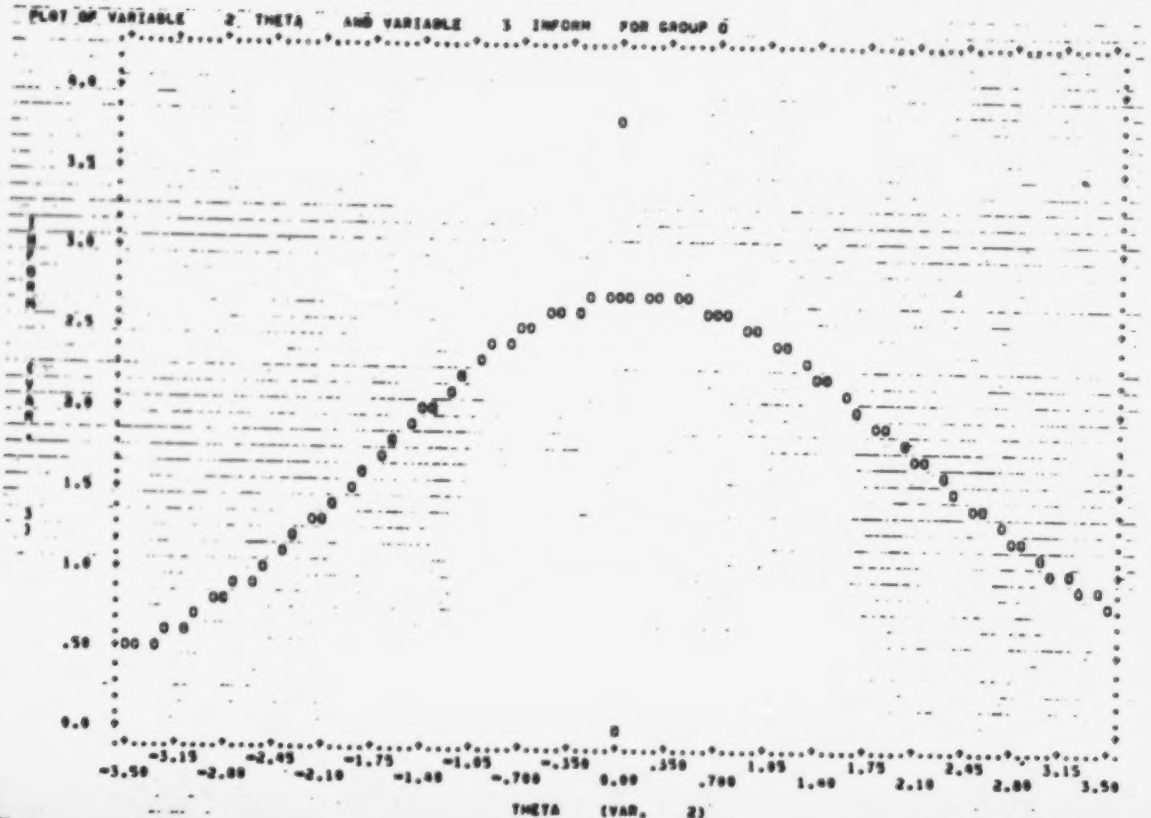


BEST COPY AVAILABLE

open-ended questions



summary



In this paper I will discuss the construction of some objective based tests for reading comprehension (mother tongue) applying the Rasch model. These tests aim to measure the ability level of students from the first three forms of the several types of secondary schooling, vocational as well as non-vocational, existing in The Netherlands. The students are between 12 and 15 years old. I will focus mainly on the construction proper, but before doing so I shall consider briefly some aspects of the theory of reading comprehension and the reasons for using the Rasch model.

It is only feasible to construct objective based tests for reading comprehension if one assumes that reading comprehension is a set of different skills. The theory is not all too clear about the truth of that assumption. In fact, there is an ongoing debate between those who support this notion and those who assert that reading comprehension is a single global skill not very different from the skill of reasoning. There has been some empirical research intended to try and resolve this problem, but it has not been conclusive. I mention Davis (1944, 1968), Spearrit (1972), Chapman (1971), and Andrich and Godfrey (1979) as advocates of the different skills theory, and Thurstone (1946) and Thorndike (1973) as supporters of the global skill theory. Whether one or the other theory is the right one is not altogether an academic question. There are some very important implications for the teaching of reading comprehension. If the different skills theory is true, teachers will have to plan carefully the teaching of each skill; if the global skill theory is true, there is no justification for teaching different skills.

As long as theory offers no solution, it seems safest for a teacher to adopt the different skills theory. At least no harm can come from that. And if he does so, objective based tests can be very helpful.

Now the tests we have in mind will not only be objective based, they will also be level tests. It is assumed that the different reading comprehension skills are such that it is unlikely that pupils of 12 - 15 years of age have mastered them completely. They reach a certain ability level and improve on it, but presumably never reach full proficiency.

Bearing this in mind it would be nice if we could construct an instrument with which it would be possible to measure a person's level of performance on a well-defined reading comprehension skill as well as his improvement on that skill.

What we want to measure, however, is not a visible characteristic such as height or weight. It is a mental ability defying direct observation. Such an ability is a latent one. It can, however, reflect itself in observable behaviour. In fact this is just what happens when a

person takes a test. His score on this test is an observable reflection of some latent ability such that we can interprete it as a manifest indicator of that latent ability. To elicit the level of this latent trait from the indicator we need a model that specifies the relation between both. Such a model for instance is the well-known classical test theory. For our purpose however this is not as suitable as we would wish it to be. There are several reasons for this, of which I mention the two more important ones.

(1) In the classical model the testees' scores are test dependent. This entails that we don't know the ability level of the testee as such but only the relation of his ability level to the ability levels of the other testees that took the same test.
(2) The item parameters in the classical test theory are not sample free. So we don't know the difficulty of a specific item as such, but only the difficulty for the sample of testees that took it.

With the Rasch model we don't have these problems: the ability levels of testees are test free and the item difficulties are sample free. It should be clear from this that the Rasch model has some real advantages over the classical model.

One of the more important assumptions in the Rasch model is that all item responses reflect a single latent trait. Now if this should not be true the model is not applicable, so it is absolutely necessary to check on it. To see whether the model fits our particular application, we use the graphical method described in Rasch (1960). I will return to this later on.

When constructing objective based tests, naturally the first thing to do is to collect a set of relevant objectives. And as it is up to the teachers in the field which objectives should be tested, the obvious thing to do is to ask them. This we did and we ended up with a prelimenary set of five objectives:

(1) the student has to identify the main topic of a passage/text;
(2) the student has to identify the main argument of a passage/text;
(3) the student has to identify conclusions;
(4) the student has to identify causes;
(5) the student has to identify effects.

The construction of tests for these objectives was started in 1980. There are none available yet; the first four will be issued in the spring of 1984.

As the ability levels of the students for which the tests are designed vary widely and as we strive at a relatively high precision in the measurement of each student's ability, we need a rather large pool of homogeneous items. In 1980, when we started, we estimated the number of items needed for one test to be about one hundred and at this moment we estimate it to be about ninety.

Each item consists of a passage or text, the length of which is at most some 250 words, and precisely one multiple choice question with four alternatives. These items are constructed by a group of proficient teachers in the field, who have a sound knowledge of at least one of the relevant types of schooling in which the tests will be used. They construct 125 items for each test, as we take into account that some items may turn out to be

bad or misfitting the model. During the construction period all items are thoroughly screened. With the 125 items that in the end are considered to be good items, an initial pretest is run. This is a very complex matter, since we have so many items and such a large range of abilities. Because of the number of items we can not ask each student to answer them all, since it would take too much time. And because of the large range of abilities, we must make sure that the sample - as we can not take the whole population - is such that the estimation of item difficulties is possible on all levels. The sample we use consists of nine groups: the first three forms of each of three types of secondary schooling. We try to get at least 150 students in each group. Each of them is asked to answer 25 items. Ten of these are common items (anchor items), which are used to equate all item difficulties, that is, to put them all on the same scale. The item difficulties are estimated with the help of a Rasch analysis.

The first pretest aims to discover which items are good ones and which are not. We check two things. The student's responses are used to see whether there is anything wrong with the alternatives. Of course we do not need Rasch analysis for this. It is nothing new. However, in addition to this we check if the items fit the model. As I have mentioned before, for this we use the graphical model test described by Rasch (1960).

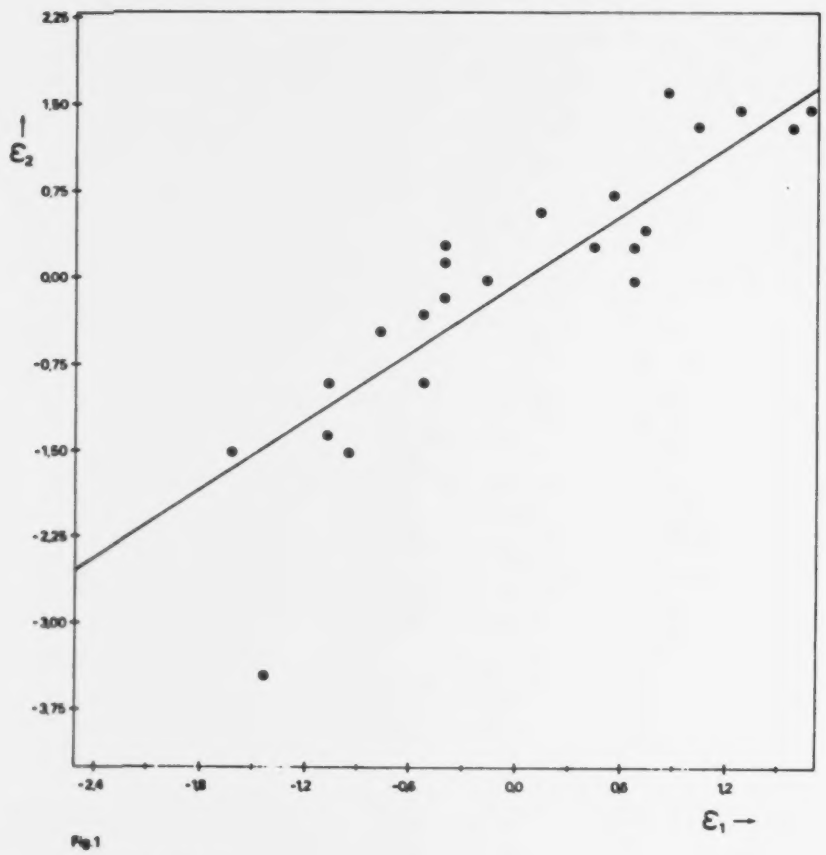


Fig 1

Graphical model fit

Figure 1 gives an example. It shows the testing for model fit of 25 items taken by one of the nine sample groups. The test is based on the assumption of the Rasch model that the item difficulties are sample free. To see whether this is true the sample group is split in two halves. For each half group the item difficulties are calculated. The item difficulties of one half (E_1) are set out on the abscis and of the other (E_2) on the ordinate. Then the items are plotted on the graph according to their two difficulties. In the ideal situation these difficulties should be the same and so the difficulties of the 25 items in the plot should lay on a straight line at an angle of 45°. In reality they appear scattered along this line. As long as the items are relatively close to the 45° line, there is a fair chance that they are unidimensional, i.e. that they measure the same latent trait. It should be stressed, however, that there is no exact criterion for when the item is so far away from this line that it can be said that it does not measure the same trait as the other items. Only one thing is certain, and that is that the further away it is, the less likely it is that it measures the same trait. In using the model test we have to decide what distance from the ideal line we consider to form a convenient boundary. In the case represented by figure 1 we discarded the item in the lower left hand corner and used the others since they fitted the model better.

After the first pretest we discard all items that do not fit the model. Moreover, faulty items are corrected if the faults are minor ones, but otherwise they are also discarded. As we have constructed 125 items and we only want to save 100 for the test itself, we can throw away 25. Up till now we have not encountered any problems in getting 100 good items.

In the last construction stage the items have to be ordered according to difficulty. To achieve this we put them through a second pretest, the sample characteristics and design of which are the same as those of the first one. With the help of a Rasch analysis we calculate the item difficulties and order the items according to them. The calibrated and ordered items constitute the measuring instrument. It is at this moment not yet decided in what form precisely the items will be offered to the teachers. Probably it will be best to order them in branched tests. In that case students will only have to take a small subset of the items to have their ability level estimated. It seems to be the most economical way to use this kind of test.

In a couple of months I hope to be able to issue some more information.

Andrich, D. and Godfrey, J.R. (1979) Hierarchies in the skills of Davis' Reading Comprehension Test, Form D: an empirical investigation using a latent trait model, *Reading Research Quarterly*, 14, 182-200.

Chapman, C.A. (1979) *A test of a hierarchical theory of reading*. The University of Chicago Press. Chicago.

Davis, F.B. (1944) Fundamental factors of comprehension in reading, *Psychometrika* 9, 185-197

Davis, F.B. (1968) Research in comprehension in reading, *Reading Research Quarterly* 3, 499-545

Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*, Danmarks Paedagogiske Institut, Kopenhagen.

Spearrit, D. (1972) Identification of subskills of reading comprehension by maximum likelihood factor analysis, *Reading Research Quarterly* 8, 92-111

Thorndike, R. (1974) Reading as reasoning, *Reading Research Quarterly* 9, 135-147

Thurstone, L.L. (1946) Note on a reanalysis of Davis' Reading Tests, *Psychometrika* 11, 185-188

There are no final examinations in the ordinary sense of that phrase in Sweden. Standardized tests are given, however, both in the comprehensive school and in the upper secondary school. These tests are usually given more than a year before the students leave school. The main objective of the tests is to help teachers get an idea of the overall standards of their classes rather than grade individual students. In this paper I will describe the tests of English, French, and German. I will try to describe the traditional format of these nationwide tests, the problems we are facing and the new ideas that are being introduced.

Background

The Swedish School System.

Swedish children start compulsory schooling at the age of seven. There is a nine-year comprehensive school, divided into three levels: lower, middle, and upper, each of three years. Each age cohort consists of 100,000 - 120,000 pupils. There are thus about one million children of compulsory school age.

At the age of sixteen, the pupils can leave school or enter *gymnasieskolan* (the upper secondary school). There are five different three-year streams and sixteen two-year streams in this school. About 40,000 pupils enter a three-year stream in preparation for further academic studies. More than 60,000 go into the two-year streams, some of which prepare the pupils for higher education (teacher training, etc), the others for various vocations. Between 80% and 90% of the sixteen-year-olds enter *gymnasieskolan* (in some parts of the country almost 100%).

Foreign Language Study.

English is studied as the first foreign language by all Swedish children. It is introduced in grade three (but more and more schools are going back to a later start in grade four) and is compulsory for the remaining seven years. On finishing the *grundskolan*, the student will have had between 400 and 500 hours of English. In the upper secondary school, most students have another 200 hours of English.

During the last three years of the comprehensive school the students can choose (among other things) German or French. The second foreign language, known as 'language B', is studied for a total of approx. 250 hours and is taken by 60-65% of the pupils. In the upper secondary school a third language, referred to as 'language C',

can be studied. This can be German or French, or it can be Spanish, Russian, Italian or Finnish.

The NBE.

The National Board of Education is a government office in charge of all Swedish schools. It works independently of the Ministry of Education. The NBE produces the nationwide curricula for both the comprehensive school and the upper secondary school which all schools have to adhere to. The overall principles and the main goals and objectives of the curricula are law, passed by parliament. The NBE also has a supervisory role in relation to the schools. One of the tasks of the NBE is the production of standardized tests given in order to help teachers grade their students and also in order to indicate to teachers and students the main aims of the curricula. The tests are thus given by a state agency on a nationwide basis in order to produce correct and comparable grading in the whole country.

Curricula.

The nationwide curriculum valid for the comprehensive school, known as *Lgr 69 (Läroplan för grundskolan 1969, 'Curriculum for the comprehensive school, 1969')*, has been used in grades 1 - 9 in all schools in the country for more than a decade. A new curriculum, *Lgr 80*, was passed by parliament in 1980 and will take effect in the autumn term, 1982. The curriculum lays down what subjects must be taught, the number of periods designed for each subject, the goals and objectives to be achieved in each subject, the principles of grading, etc. The curriculum is a book of about 200 pages.

A similar curriculum, *Lgy 70, (Läroplan för gymnasieskolan 1970, 'Curriculum for the upper secondary school, 1970')*, is used in all upper secondary schools, i.e. for grades 10-12. This curriculum will be revised in 1985 when the whole upper secondary educational system is expected to be reorganized.

Grading

Grades are given in the comprehensive school in grades 7, 8 and 9 as well as in the upper secondary school. All grades are norm-referenced with grades 1-5 given to 7-24-38-24-7% of the pupils respectively. This 'normal distribution' of grades applied to the whole country, not the individual class. In order to give just and correct grades in her class, the individual teacher needs some help in determining what the standard of her class is in relation to the country as a whole. This help is provided by the standardized tests.

Standardized tests

There are two different kinds of standardized test. One is known as *standardprov*. These are given in the comprehensive school and are not compulsory. The other kind is known as *centrala prov*. They are compulsory. These tests will, for the sake of brevity,

henceforth be referred to as SP and CP respectively.

Aims

As has already been mentioned, the main aim of the standardized tests is to help teachers give correct grades. There is no pass/fail limit: grade 1 and grade 2 are not considered failing grades. The tests also have the effect of influencing teachers and students. This 'wash-back effect' is intentional and important. The inclusion of listening comprehension tests has certainly raised the standards in this area. The lack of a speaking test and of a written production test may very well have influenced standards adversely in these areas. The tests are also used, to some limited extent, to evaluate the standards in the country.

SP and CP Tests

Standardized tests in foreign languages, SP, are given in the comprehensive school as follows:

English in grade 6, after four years of study; this test consists of four subtests, it is almost ten years old and will not be renewed; the test is confidential but can be ordered by researchers etc.

in grade 8, after six years of study or 400-450 hours; there are two versions, each of 150 items, overlapping by about 50 items; there are four subtests (Gramm & Voc 1 and 2, RCT, LCT); the tests given in 1982 are about ten years old and will not - hopefully - be used again: new tests are being pretested this year and will be introduced in 1983; these tests are confidential in principle.

French in grade 9, after three years of study or about 200 hours; there are three subtests (Gramm & Voc, RCT,LCT); this test has been used for several years but will not be renewed, in future no standardized test will be given (unless present decisions are changed); confidential but can be ordered.

German in grade 9, same as for French even though some of the techniques used are not the same; in both French and German there are two versions of the tests as in English, one easier and one more advanced, partly overlapping.

The SP are not compulsory but the vast majority of all teachers use them (except, lately, English grade 6). The English tests are thus taken by about 100,000 pupils every year.

In the upper secondary school there are nationwide, standardized tests, called *Centrala prov*, CP, in English and in French and German as 'language B' (as well as in several other subjects). There is no such test for

'language C' students. The tests are given in the Spring Term in grade 2 of the upper secondary school ('grade 11'), i.e. after eight and a half years of English and four and a half of French or German, or a total of approx. 500-600 hours and 300-350 hours respectively. These tests differ from the SP in that new tests are produced every year and in that the tests are compulsory. There are the following tests:

English in grade 2^3 i.e. the second year of the three-year courses; there are several different subtests: Gramm & Voc A and B, RCT, LCT and a cloze-like integrative test; new tests are produced every year and the old ones used in schools for practice purposes after they have been given.

English in grade 2^2 i.e. the second year of the two-year courses; there are two versions of this test, one easier and one more advanced with about 30% overlap; these tests are confidential and cannot be used for practice, parts of the tests have been used more than once but the full test is not repeated.

French in grade 2^3 with four subtests: Gramm & Voc, Dictation, RCT, LCT; some of the techniques are different from those used in English; new tests are produced every year and the old ones used for pratice purposes.

German in grade 2^3 with three parts: Gramm & Voc, RCT, LCT; experiments with free written production have been carried out in German and, to some extent, in French; models for the testing of oral production haven also been produced in French and in German.

Test Construction

The tests are constructed, under the auspices of the NBE, by groups of experts consisting of psychometricians, language teachers, NBE inspectors, NBE FL consultants and native speakers. Texts, etc. are written and/or selected and items constructed in 'year 1', several try-out versions are pretested in the spring of 'year 2', analysed by computer and final versions put together in the autumn and then the tests are given in the spring term of 'year 3'. A large sample of tests is collected and analysed, grade limits are laid down and further analyses made in the autumn of 'year 3'. Within the GUME/ Tests Project Group we have taken an active part in the construction of the English tests, carried out item analyses of both open-ended and multiple-choice items in all three languages, produced the grading norms and, to some extent, made further analyses of the test results.

For pretesting purposes we use students in grade 11, i.e. those who take the regular test. This means that certain comparison can be made between the regular test and the new one. The students who take the try-out versions are

thus one year ahead of those who take the actual test. Between 200 and 400 students take each try-out version. Many more items are pretested than can actually be used. Item analysis data in the form of facility indices and discrimination indices (point-biserial correlations) are calculated and used in the selection of items. Both teachers and students taking part know that a new CP is being tried out but we have never experienced any problems with the confidentiality of the tests.

Test Administration

The CP tests in grade 2[3] are given on certain days, the same for each individual language all over the country. The other CP and all SP tests are given within a three-week period. These latter tests, which are confidential and have been given in more or less identical form for several years, can thus be given on any day within this period. The fact that the tests are not final examinations but rather given to guide the teacher in grading the whole class probably explains why the tests can be given within a period rather than on one particular day and why the same test can be repeated several years in succession.

The tests are marked by the teachers themselves according to a pre-prepared key sent to them together with the tests. Most items are marked dichotomously right or wrong (in French, one point is sometimes awarded for an acceptable and two points for a completely correct answer) and the number of correct answers added up. Sample of students' booklets (e.g. those for pupils born on the 3rd, 13th and 23rd of any month) are collected and grading norms established. New norms are laid down every year for the SP tests. The grades are thus strictly norm-referenced and not criterion-referenced. Grade 5 does not, therefore, represent any absolute measure of language proficiency. However, despite the fact that several thousand students take each test (in the case of the English SP around 100,000) the standards differ very little from one year to the next.

Content and Format

The standardized tests, both SP and CP, have a history that goes back to the 'psychometric-structuralist sixties'. The tests are 'itemized', and each item is marked as correct or incorrect. There is a stress on reliability and comparability. Therefore the multiple-choice format has been widely used. Usually, there have been four parts: grammar, vocabulary, reading comprehension and listening comprehension (and, in French, dictation). Written production has been kept to a minimum of one word or just a few words. There are, however, experiments under way both with free written production and with oral tests. In all languages open-ended items are used but these are also usually scored right or wrong. Longer pieces of text or essays have not been part of the present tests used for grading purposes. The Grammar-Vocabulary-Phraseology parts are increasingly 'integrative' and, in a way, quite similar to cloze-tests. To answer an item correctly, reading comprehension

is demanded as is an understanding of 'wider contexts' and an 'expectancy grammar' (Oller: Language Tests at School, 1979; p.16 ff.). A cloze-like test, sometimes referred to as 'cluze', with pretested gaps at suitable points in the text (rather than a mechanical deletion of every n:th word) has also been used lately.

In the last few years there has been an increasing stress on validity, i.e. on the goals and objectives of the curriculum and the extent to which these are reflected in the tests. Because of this the number of multiple-choice items has been reduced and open-ended items introduced to an increasing degree. Various techniques to test ordinary written production have also been tried out. These will be described below. This stress on validity rather than on psychometric and statistical efficiency explains both why the listening-comprehension test has been considered essential and why so many teachers and test constructors hesitate to use the pure cloze-test.

Since the tests are used compulsorily by all teachers and for all students in all schools in the country, changes in test content and format must be introduced slowly and preceded by careful information. New kinds of tests are tried out for a number of years before they are introduced in the regular SP and CP tests.

Tests of elements and skills

Grammar-Vocabulary Tests

Even though the objective of foreign language study is the achievement of 'the four skills', there has always, by tradition, been a strong stress on the 'linguistic elements'. There are still in all standardized tests grammar and vocabulary items. No strict limit is drawn between grammar and vocabulary tasks and the formats used are almost always such that the items also test reading comprehension. If, in the multiple-choice format, the context used is more difficult than the word or phrase tested, the item is considered a reading comprehension task even though it may look, on the surface, exactly like a vocabulary item.

An example of a traditional-looking item is given in example 1, taken from this year's CP test in English.

(1) It's rather less _____ to call a girl 'plump' than to call her 'fat'.

A complimentary
B rude
C flattering
D indecent
E polite

The correct answer was chosen by 60% of the students who took this item in the pretesting sessions last year. The point-biserial correlation is .32. All the distractors were chosen: polite by 22%, r_{pbis} -.23, and flattering by 10%.

An example of an open-ended item which proved to work well is given in (2).

(2) I enjoy cooking, but sometimes it's fun _____ a meal even if it is more expensive.

The students were told that two or three words were missing, never more than five. Many correct answers were possible. About 55% of the students produced 'to go out for' or some similar phrase. There were 113 different answers; 4% of the students were not able to suggest any answer at all. This kind of item is well liked by the teachers and usually shows high discrimination indices. There are some problems in using this technique among low-proficiency students: too many variants are produced and the number of spelling mistakes is usually extremely high.

Some of the items found under the 'Gramm & Voc' heading are of a communicative kind. Even though the multiple-choice format is used, the items can test what has come to be called 'linguistic functions'. In example (3), which is taken from the German SP test for low-proficiency students in grade 8, the student has to select the best answer to the question or phrase given in a picture.

(3)

A Das konnte ich nicht.
B Das wusste ich nicht.
C Das habe ich nicht gebracht.
D Das habe ich nicht gesucht.

In example (4), taken from the French SP test, the student is supposed to select the best question to fit the answer given in the picture.

(4)

A Vous achetez souvent ce journal?
B Est-ce qu'elle achète le journal?
C Quand achetez-vous ce journal?
D Vous achetez souvent des revues?

Both these items seem to prove, in my opinion, that the multiple-choice format can be used for a kind of communicative test: old techniques can be used for new purposes. Those readers interested in studying further examples from the CP tests are referred to the book *Swedish Foreign-Language Standardized Tests - a collection of examples of Centrala Prov in English, French and German with short comments* by Torsten Lindblad (ed.), or to the actual tests themselves.

Reading Comprehension Tests

The reading comprehension test used in the CP tests usually consists of a prose passage broken into several sections, each followed by a number of questions. The main aim here is to test not only the students' ability to interpret the salient points correctly but also to draw conclusions from implications within the text. In constructing this type of test we actively avoid questions based on minor details. We try to produce questions testing both 'atomistic' reading, i.e. the exact understanding of the text, and of 'holistic' reading, i.e. the understanding of overall ideas. The latter kind of question, however, often proves to be the more difficult to produce. These RC tests are fairly similar in the three languages and always of the multiple-choice kind. We have tried using open-ended answers but the variety of responses makes is difficult to produce a scoring key that can be used by thousands of teachers.

A new kind of RC item has been introduced recently. It consists of a mini-text from which a word or group of words has been omitted. The difficulty for the students lies in the interpretation of a fairly complex text.

(5) Mountaineering in this area is often _____ because of blizzards and avalanches.

A dangerous
B recommended
C great fun
D practised
E exciting

The alternatives given are so linguistically simple as to present no difficulties to students at this level.

Listening Comprehension Tests

It is difficult to illustrate listening comprehension tests in writing. The interested reader is therefore referred to the tapes that accompany the various tests and which can also be ordered free of charge. The following general remarks can be made. The texts of the LC tests usually consist of dialogues, discussions or uninterrupted narratives (a person tells a fairly long story or reads out something). Educated or everyday spoken language is used. We have not successfully used recordings of actual dialogues in the street or radio broadcasts. Usually the texts have been tampered with in order to facilitate the construction of distractors. The LC tests are always in the multiple-choice format. Students and, often even more, teachers usually react strongly against LC tests where dialect or non-British variants are used or where there is background noise.

Even a moderate American or Australian accent in an English test is usually commented on.

One problem in the construction of LC tests in German is that they tend to be too easy for Swedish students. The pupils are usually supposed to show that they understand the overall ideas and outlines of the text they listen to. It is the opinion of the test constructors that the difficulty of the tasks should be the text itself, not the questions. They sometimes deviate from this rule, however, to attain an adequate degree of difficulty in the tests.

Over the years attempts have been made, in various ways, to minimize the memory factor when constructing LC tests. In earlier tests, for instance, the questions were presented before the pupils listened to the text. They knew, then, in advance what information to look for, which is usually the case in real life. As a result, however, the pupils sometimes did too well from the tester's point of view. That is why this method has been abandoned for the time being. Questions are now asked after every passage only, in German as well as in English and French.

French is, of course, the most difficult of the three foreign languages for Swedish students. It is interesting to note that it is the French CP tests which have used actual real-life dialogues to the greatest extent. A large amount of material has been gathered in English, however, and there is work going on to produce similar tests in the first foreign language also.

The LC tests that we use usually show the same reliabilities and correlations as the other subtests. The reliability is often .85, sometimes even higher. From a 'functional' or 'communicative' point of view, however, they may not be very valid. This is probably the area in which we have not been able to keep up with the international development as well as we ought to. We do find, however, that from the testing point of view our LC tests work extremely well and that they serve their purpose of influencing the teaching in our schools in a desirable way. It is doubtful whether the introduction of more lifelike LC tests would bring about very dramatic changes in teaching or learning.

Written Production

As has been mentioned above, written production is only tested in open-ended items in which the student is supposed to supply one word or a maximum of 5-6-word wordstrings. Many different techniques are used to prompt the students to produce these words. It should be stressed that translation is never used.

Experiments are being carried out with various kinds of 'free production' or guided composition in order to make the students produce longer passages of text. Various techniques are used. The student can, for instance, be given a short introduction in the mother tongue and then told to write a letter or describe a scene in a picture.

A number of key words are given and should be included. The greater problem in testing written production is the marking or grading of the students' essays. Usually a larger number of students' scripts have been collected, read by a group of experts, evaluated as 'above average', 'average', or 'below average'. The teachers are then given a number of examples of each category and can compare their own students' scripts with these models. Some examples are given in the previously mentioned book *Swedish Foreign-Language Standardized Tests* (see also list of references). There is great disagreement among teachers and test constructors as to what stress should be put on linguistic correctness as compared to communicative content.

One further problem which has not been solved is the following: How is the teacher to arrive at one grading of each student if she has a sum of, say, 80 points on an objectively scored test and the result of a written production test on a 3-point scale? It is difficult and complicated to add together results from a reliable, itemized test and a more subjectively scored productive test. The teacher will have to solve this problem somehow in giving her own grade to the students, but if the test results are to be used to produce nationwide norms, the problem becomes much greater.

Oral Production

The students' oral proficiency is not tested in the standardized tests. The teachers are supposed to evaluate this regularly under normal teaching conditions and to give it its proper weight in grading the students. There is a risk, however, that the students' actual proficiency is either underestimated or overestimated due to various factors. It is therefore considered a good idea to try and test the students formally some time during the school year. Formerly this was done in the oral final examinations. Since there is no such examination any longer, there is a need for some kind of testing procedure which can be used in addition to the standardized tests.

There are two problems in testing students' oral proficiency: How do we motivate the students to say something and how do we evaluate what is said? Various techniques have been tried, and within the GUME/Tests Project Group we have been working on this in a number of different languages for different student categories. In the context of CP testing, the French group has produced most material. Two examples from this will be given (see also list of references). The student works with two tape-recorders, listening to instructions and parts of dialogues from the one and recording her answers on the other. To do the task in example (7) the student needs two pictures.

From 'tape 1' the student first hears a waiter asking the question, in French, whether she wants an ice-cream. The student is instructed in the student's booklet in Swedish to reject that offer and ask for a banana instead. There are several such items and the student knows that the

first picture refers to the question asked, which should be reacted to, and the second picture to what the student should say.

(7)

In item (8) the picture serves as a very simple prompt to the context. In the student's booklet the student is instructed in the following way: "You are sitting in a train and have nothing to read. In front of you there is a man who has several newspapers ..." This is a communicative kind of test. The student is supposed to show whether she can fulfil the 'function' of 'asking for something'. The previous example was supposed to test 'giving information'.

(8)

There are also carefully worked out instructions as to how various answers should be graded. In the case of the French examples just given, each answer is awarded a number of points according to a table of figures for correctness and communicative content. In other contexts, descriptions of levels of proficiency are given for pronunciation, vocabulary, grammar, and fluency and the evaluation made on the basis of a number of answers.

Problems in SP and CP Work

There are many problems involved in the production of nationwide large-scale tests which influence students' grades and thus their whole future. I would like to end this paper by mentioning four problem areas that we are discussing in Sweden at the moment.

1 How do we combine demands for high reliability and 'objective' scoring with high-validity 'communication-oriented' techniques?

2 Is there a risk that we are coming full circle back to the 'prescientific' model (Spolsky)? There is, in my opinion, a risk that in going over from the structuralist-psychometric era to a period when concepts like validity and communication are stressed, we give up all statistical demands and rely on the omniscient examiner who can evaluate a student quickly in an oral examination and with the help of an essay. Perhaps some kind of eclecticism is the answer?

3 How do we balance reasonable demands for correctness (spelling, grammar) and communication-oriented tasks? How we try to do this in Sweden at the moment has been described above.

4 Can the same test fulfil many functions, e.g. for grading and evaluation? Is it, for example, 'permissible' to use student answers to a test given for norm-referenced grading purposes to make qualitative analyses and to use the results diagnostically? In a paper read at the 4th International Language Testing Symposium in Colchester (Lindblad, 1982), I described a computer technique developed in Gothenburg to analyse student-produced answers to open-ended items which can be used for this purpose.

The Future

In the future there will not be any SP tests in French and German - unless present NBE decisions are changed - but the English SP test will be renewed every three years. Hopefully a new one will be ready in 1983. The three CP tests will all continue to be produced and the experiments with new models and with tests of oral and written production will be expanded and followed up. New kinds of listening comprehension tests are also being tried out.

The National Board of Education has been completely reorganized as from July 1, 1982, and the Department of Test Construction and Evaluation will disappear. The responsibility for producing the tests in English (and, probably, later in French and German) will be taken over by the Department of Educational Research at Gothenburg University. Officially, the NBE will still be responsible for all standardized tests, however.

Now that we are setting out on the challenging task of building up a kind of 'mini-Cito' for foreign-language testing in Gothenburg, we hope to be able to establish good international relationships. I hope this little paper on the Swedish standardized tests can serve as a kind of invitation to an exchange of ideas, and of actual tests, and to a discussion of results and new techniques!

Lindblad, Torsten (ed.) (1981) *Swedish Foreign-Language Standardized Tests* - a collection of examples of Centrala Prov in English, French and German with short comments. The GUME/Tests Project Group, Department of Educational Research, Gothenburg University, June 1981.

Lindblad, Torsten (1982) 'Computer-based Analysis of Open-ended Foreign-Language Test Items'. Paper presented at the Fourth International Language Testing Symposium of the IUS, Colchester, September 1981. In Culhane, T. et al. (1982) (eds) *Practice and Problems in Language Testing*, University of Essex, Dept. of Language and Linguistics, Occasional Papers No. 26.

Några modeller för skriftlig produktion (Some models for written production), häfte 1, 2 and 3 (three booklets of about 50 pages in all) published in Spring 1979, for French in grade 3^3: Book 1: production of sentences or parts of sentences; Book 2: production of texts; Book 3: examples of expected answers and dialogue manuscripts.

Test för bedömning av muntlig produktion i franska (Test for the evaluation of oral production in French) grade 3^3, French as the second foreign language, i.e. after six years of study, produced in February 1982. There are two regular tests and two for practice purposes, each with a pre-recorded tape; the student has to use two tape-recorders, one to listen to instructions, one to record her/his responses.

Wie schreiben schwedische Jungen und Mädchen Deutsch? Untersuchung eines Versuchstestes 'gelenkter freier Produktion' by Lil Danielsson; this is a master's thesis written at the German Department, Stockholm University, Spring 1981.

Das Zertifikat Deutsch als Fremdsprache - Erprobung und Auswertung an schwedischen Gymnasien im Frühjahr 1978, 80 pp, produced by the NBE in cooperation with the Goethe Institut, Stockholm; describes tests of both oral and written production.

Addendum

Any researcher or professional test administrator interested in having any of the tests described in this paper (including tapes for the LC tests) can order them, free of charge, through me (Gothenburg University, S - 431 26 Mölndal, Sweden). We would welcome, in return, information on the way they are used as well as data and comments.

Language Testing at about Age 16:
Current Developments in England,
Wales and Northern Ireland,
Robert H. Osborne

Before I begin, may I first make a few remarks about the organisation of education in the United Kingdom and how this affects what I shall be talking about. I hope you'll forgive me if I'm covering ground that you're already familiar with.

Much of the responsibility for education rests with local authorities which are part of the local government structure of town and county councils. But the system is controlled, over all, by central government and Acts of Parliament. The system operating in Scotland is significantly different from that which operates in the rest of the United Kingdom and I shall not be referring to developments there. But what I have to say is relevant not only to England but also to Wales and to a certain extent Northern Ireland as well.

What I have to say is in two parts. First, I shall be talking about a survey of language performance in schools and then I shall outline developments towards a new system of examining to replace the existing GCE 'O' level and CSE examinations. These two projects I see as having, potentially, a great beneficial effect on teaching and will also have an effect on the type and range of language testing at about the age of 16 in the future.

Language Performance in Schools

In 1975 a unit was set up in the Department of Education and Science - the government's education office for England - to provide information about general levels of performance of children and young people at school and how these change over the years. The unit is called 'The assessment of Performance Unit' or, more commonly, the A.P.U.

Its terms of reference are:

'to promote the development of methods of assessing and monitoring the achievements of children at school, and to seek to identify the incidence of under-achievement.'

and as a result it has undertaken a research programme which is defined in terms of four specific tasks:

Firstly it has sought to 'identify and appraise existing instruments and methods of assessment' insofar as they may be relevant; secondly it has undertaken 'the creation of new instruments and techniques for assessment'; thirdly, it is carrying out assessment in co-operation with local education authorities and schools; and fourthly it has sought to 'identify significant differences of achievement related to the circumstances in which children learn, including the incidence of under-achievement, and to make the findings available to those concerned with resource allocation within government deparments, local education authorities and

schools.'

Before I go any further I must make it clear that this is seen as a long term project and although reports on language performance at ages 11 and 15 have been published, it is intended that these should be only the first of a series which will span a number of years.1 For 15 year olds the aim is to carry out five national surveys initially and no doubt each will be the subject of a report. And as well as being long term, the project is also very broad in scope and will try to recognise language performance in parts of the curriculum which do not receive the direct impact of language teachers. Maths and Science are being treated in the same way and a certain amount of exploratory work has also been undertaken with regard to aesthetic subjects. One other point; all the monitoring of language performance that has taken place so far has been concerned with English, but monitoring of French, German and Spanish as first foreign languages is planned to begin in 1983. Language monitoring of 15 year olds began in 1979 using assessment procedures that had been developed by a team based at the National Foundation for Educational Research. Of course, national surveys of attainment are not new and it was in fact the NFER that was responsible for surveys of national attainment in reading at the ages of 11 and 15 which were conducted in the period 1948 to 1964. The reading materials developed for the APU surveys are different but it is interesting to note that some of the outcomes of the initial monitoring can be linked to the results of the earlier reading surveys. More of this in a moment.

First some figures to indicate the scale of the survey. A representative national sample of about 9,100 pupils from 378 schools, - state schools, maintained schools and independent schools - were involved in the survey. Approximately 5,900 of the pupils were from schools in England, 1,400 from Wales and 1,800 from Northern Ireland. However, only reading and writing skills were assessed and only seven hundred of the pupils completed tests which linked reading and writing. Of the remainder, 5,300 pupils completed tests in reading and 3,100 completed tests in writing.

The tests in reading were based on eight illustrated booklets varying in content, structure, source, style and purpose. In the main, the pupils had only one booklet to deal with but four additional booklets were produced which contained material from more than one of the other booklets and these were then used with a sub-sample of the pupils. This, as you will have realised, was to ensure that the test items could be scaled, and to help further with this, all the pupils were asked to read one set passage and answer fifteen questions related to it.

The report examines, in great detail, the pupils' responses, with regard to the short story and travel brochure, but the information that has emerged will really only become significant when further surveys have been completed. Of more interest is the fact that nearly a thousand of the pupils who took the reading test also

took one of the earlier national reading tests - National Survey Test 6 - which had been in use since 1955. There was no statistically significant difference between the mean scores obtained in 1970 - 71 and 1979, and the correlation between the 1979 NS6 and APU reading tests was 0.81.
A sub-sample of pupils who took part in the reading survey also completed a questionnaire about their reading preferences and interests. The main outcomes were firstly that boys were significantly less favourably inclined than girls to reading, both in and out of school. Secondly, there was a consistent but low correlation between attitude and attainment in reading. Thirdly, the compulsory study of a narrow range of books in school appeared to encourage a dislike of reading and fourthly, when the results of the survey of eleven year olds is taken into account, the enquiry shows a general decline in enthusiasm towards reading from the age of eleven.

Turning now to the work undertaken with regard to writing, ten booklets were used, each containing a writing task, an editing exercise and an additional writing task that was common to all ten booklets. As with the reading tests, the raw scores were scaled to take account of the relative difficulty of the tests. The weakest writing was analysed and the report says that about one per cent of the pupils 'provided no evidence of an ability to communicate in writing'. About nine per cent of the scripts were 'inadequate' in terms of content and appropriateness of style.

The booklets also contained three short questions on attitudes to writing, which gave approximately sixty per cent of the sample the opportunity to express a certain amount of frustration with routine written work in school. The mean scores of pupils with negative attitudes were significantly lower than those of other groups of pupils and significantly more boys than girls stated dislike of or lack of interest in writing.
These, then, have been the main outcomes of the first part of the investigation, but there are a few other matters that may be of interest. While the individual pupils' results were kept confidential, the team was able to relate them to background variables and as a result have been able to draw the following conclusions: firstly, there was no significant difference between the mean reading scores of girls and boys, but girls obtained a significantly higher writing score than boys and secondly, in England and Wales, pupils in single-sex schools achieved significantly higher scores than those in mixed schools. However, it must be pointed out that there are other factors involved, such as the type of school and the social background of its pupils, which have not been taken fully into account.

As the project continues, it will be possible to derive considerably more information from the results obtained, and the team will also be reporting on listening and speaking skills on which work has been proceeding for some time now. Also, the difference in the mean writing scores for boys and girls will be investigated further and no doubt many other matters too.

Before I turn to the system of examinations, you might be interested to know that the APU plans to deal with performance in a first foreign language by using matrix sampling techniques in order to derive a profile of performance in various aspects of language activity. As with the English project, account will be taken of background variables and if it is clear that some of the preliminary findings require further investigation it is likely that additional test material will be produced.

16+ Examinations

To those of you who are only familiar with a centrally organised system of examinations, I am sure the more I said about the British system, as it now exists, the more confusing it would appear to be. True, nearly all pupils take one of two types of examination, providing certification in individual subjects, before they leave school, the two types being one for higher levels of ability only - the General Certificate of Education - GCE - and one for a very much wider ability range - the Certificate of Secondary Education or CSE, but to go any further will indicate a tremendous diversity in approach and techniques. Nevertheless, I hope you will find some of the information I can give of real interest and in any case it must serve as a back cloth to any presentation of the developments that are currently taking place.

Firstly, then, the General Certificate of Education. There are at present eight GCE examining Boards in England and Wales and you may find that you know the names of at least some of them. (The names are listed later in this paper.)

Each of these defines its own syllabuses and methods of assessment. Some of them have been in existence for over a hundred years and were originally established to help universities to select their entry, a matriculation examination in fact. Nevertheless, they have always relied on comments from school teachers on the suitability of syllabuses and techniques of assessment and have tried to respond to changing needs. As the Oxford Delegacy said in 1857 'An examination would give a definite aim to the school masters, and a great stimulus to the scholars; and would afford an evidence to the public how far the exertions of both had been successful. It was not, however, expected or wished that the University should undertake to prescribe authoritatively any particular course of instruction, or should endeavour in any way to interfere with the system which any schoolmaster might have been pursuring. The University was asked only to test results, not to inquire into methods. The statute accordingly was so framed as to leave a wide discretion to the Delegates in regard to the subjects and the form of the examination, and also in regard to the mode in which the successful candidates should be distinguished or otherwise rewarded, and the Delegates in endeavouring to act in the spirit of the statute considered themselves bound to consult the wishes and requirements of the existing schools, and to interfere with them as little as possible.' In line with

current government policy their examinations are designed for the top twenty per cent of the ability range and largely because of traditional attitudes to such examinations the large majority are externally assessed, in other words, the person marking a candidate's work will not be asked to assess the work of his or her own pupils.

The CSE examination system was set up about fifteen years ago to cater for candidates down to the fortieth per centile. At that time children - young people - were only required to attend school until they were fifteen so the candidature in the early years was quite low. Subsequently the school leaving age was raised to sixteen and entries increased dramatically, the most recent figure for subject entries over the whole country being 3,382,334 for the 1981 examinations.2 This development has, of course, called into question the validity of trying to restrict the range catered for by public examinations. Nevertheless, the present government has shown no inclination to change its policy in this regard.

To administer CSE examinations we currently have thirteen Boards in England and Wales organised on a geographical basis, the names of most of which you may never have heard before. (The names are listed later in this paper.) In 1980, the Middlesex and Metropolitan Examining Boards merged to form the London Regional Examinations Board, and later this year the West Yorkshire & Lindsey Regional Examining Board and the Yorkshire Regional Examinations Board will join to form the Yorkshire & Humberside Regional Examinations Board.

Apart from the size of the ability range catered for, another important difference between the GCE and CSE systems is the extent of teacher control which for CSE is considerable, particularly so for my own Board, the Southern Regional Examinations Board. There is also, in SREB, a large amount of internal assessment - well over 50% - but with other Boards this tends to be rather less.

As with GCE, each CSE Board is responsible for developing its own syllabuses and examinations, and with both systems there is provision for individual schools to develop their own syllabuses but still to receive, after moderation, certification from the Board.

As you will have realised, the result of all this is a very great diversity of syllabuses and methods of assessment, no less so in language than in any other area of the curriculum. As you will see if you refer to 'Modern Languages Examinations at Sixteen Plus' one hundred syllabuses were available in 1979 for board-based examinations in French, German, Italian, Russian and Spanish, and this figure does not include syllabuses developed by individual schools and groups of schools. Cross moderation exercises are carried out regularly, of course, but it has to be admitted that while the system has the advantage of tremendous flexibility, and in the case of the GCE Boards, a certain competitive spirit, it is nevertheless somewhat unwieldy and baffles the uninitiated. In response to a growing amount of pressure from teachers, local government and employers,

particularly the main 'users' of the certificates, the government eventually decided it should do something about it and this, very briefly, is how it happened.

First they agreed to receive recommendations drawn up the Schools Council - a quasi government body, widely representative - recommendations which amounted to this: make sure that as quickly as possible the system of examining is simplified while retaining existing standards (albeit with a new grading structure) and catering for the same range of ability. The principal documents are the original recommendations made by the Schools Council, the report of the Waddell Committee (July 1978)[4] which was asked to investigate the matter and present to the government its own recommendations, and the government response, embodied in the intentions contained in its white paper 'Secondary School Examinations: A Single System at 16 Plus' of October 1978.[5]

In order to achieve a reduction in the number of syllabuses, the government proposed that groups of boards, organised on a geographical basis, co-operate in running the new system of examining. Wales and Northern Ireland already had substantial links between their GCE and CSE Boards but in England the remaining nineteen boards were asked to agree a group structure among themselves and the groups that have emerged are as follows:

NORTHERN GROUP

GCE	Joint Matriculation Board	CSE	Associated Lancashire Schools Examining Board
			North Regional Examinations Board
			North West Regional Examinations Board
			The West Yorkshire & Lindsey Regional Examining Board
			The Yorkshire Regional Examining Board

MIDLAND GROUP

GCE	Cambridge University Local Examinations Syndicate	CSE	East Midland Regional Examinations Board
	Oxford & Cambridge Schools Examinations Board		The West Midlands Examinations Board

LONDON & EAST ANGLIA GROUP

GCE	CSE
London University Entrance and School Examinations Council	East Anglian Examinations Board
	London Regional Examining Board

SOUTHERN GROUP

GCE	CSE
Oxford Delegacy of Local Examinations	Southern Regional Examinations Board
Associated Examining Board	South East Regional Examinations Board
	South Western Examinations Board

WALES

GCE	CSE
Welsh Joint Education Committee	Welsh Joint Education Committee

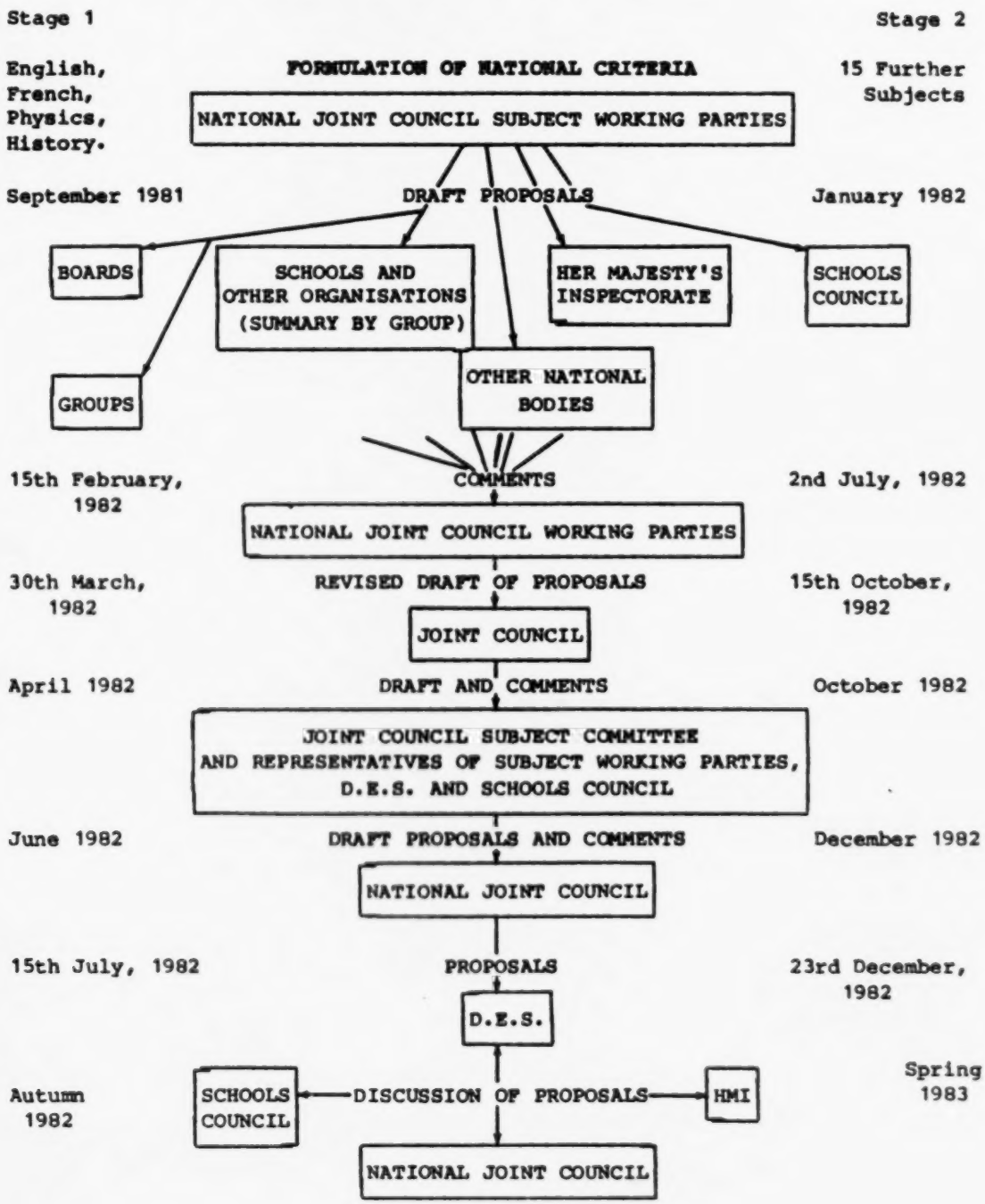
One of the recommendations of the Schools Council had been that "criteria should be established for the acceptance, validation and moderation of syllabus and schemes of assessment." (8th July, 1976)

The Government agreed that syllabuses in subjects important for subsequent stages of education or of vocational relevance have sufficient in common, and are relevant to the needs of subsequent courses of education and employment, to enable the grades awarded to be accepted with confidence by those concerned.

In order to do this it has been necessary for all the Boards to work together and for this purpose a National Joint Council was set up and this council in turn set up subject working parties and has organised a massive consultation exercise, through which it has invited teachers and other interested parties to say what they think of draft proposals. The way in which the development of criteria is being organised is shown in the following diagram:

A SINGLE SYSTEM OF EXAMINING AT 16+

General Certificate of Secondary Education?
Certificate of General Education?



Among the conference papers were the draft national criteria for English and French in the form they went to schools and other organisations. What they represent, in effect, is the extent to which Boards are, so far, prepared to define standards and the way in which examinations will be organised. The working parties have, of late, been considering the responses made by schools and other organisations and making any amendments to their proposals that they feel are necessary. With regard to French, considerable interest is being shown in graded testing and we may soon see substantial developments in this field.

With regard to English, the following points had received a surprising amount of general agreement:

1 Ability Range

The suggestion that the examination should be restricted to the top 60% of the ability range had been universally condemned.

2 Unitary English

With the exception of only three GCE Boards, it had been recommended that examinations in English Language should not be provided and that any examination entitled "English" should include some literature.

3 Orals

With the exception of only three GCE Boards, it was felt that an oral assessment should be compulsory in any examination in English as a mother tongue.

4 Course Work

There was general support for the inclusion of course work as part of any written assessment in English and English Literature.

As you will have seen, there is still a long way to go before any final decisions are taken with regard to the precise way in which the new system is designed and implemented, but I trust that the information that has been presented will help to put into perspective any of the developments that will take place in language testing, in England and Wales, at about the age of 16, over the next 10 or 20 years.

1 *Language Performance in Schools, Secondary Survey Report No. 1* published by Her Majesty's Stationery Office, 1982 (ISBN 0 11 270387 9)

2 Minutes of the Examinations Committee, The Schools Council, February, 1982 (SC 82/37)

3 *Modern Languages Examinations at Sixteen Plus, A Critical Analysis* published by the Centre for Information on Language Teaching and Research, 1980

4 *School Examinations* published by Her Majesty's Stationery Office, July 1978 (Command 7281-1)

5 *Secondary School Examinations: A Single System at 16 Plus* published by Her Majesty's Stationery Office, October 1978 (Command 7368)

Final Language Examinations in Secondary Education: the Dutch System

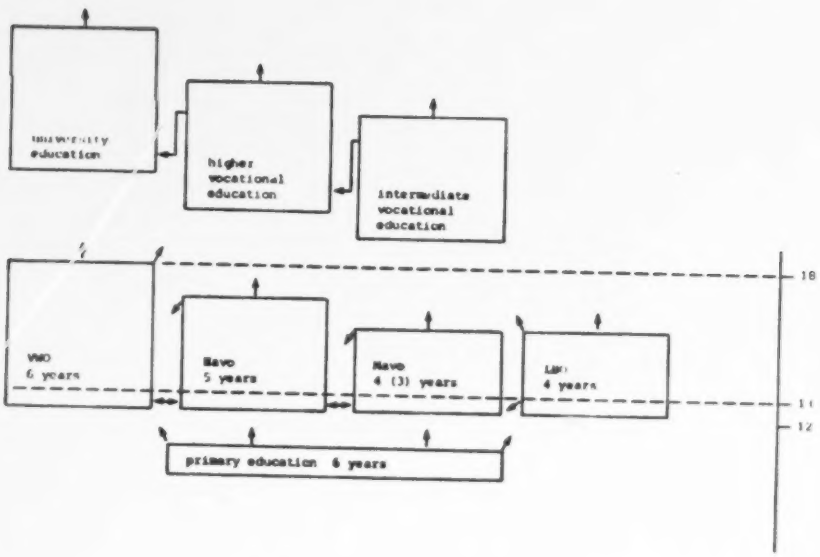
Trude A. Maas

The purpose of this paper is to give a picture of the actual situation with regard to the secondary school-leaving examinations in The Netherlands and of the proposals that have been made to change some aspects of the system. After a general view we will concentrate on the examinations of languages in particular.

The Dutch system of secondary schools

Pupils transfer at 12 plus to one of four kinds of secondary school. Who goes where is determined by the headmaster's assessment, wishes of the parents and, sometimes, an aptitude test. There are several possibilities to transfer from one type of secondary school to another if pupils develop in a different way as was expected at the age of 12. Orientation years before the final choice are common. Some secondary schools contain all four secondary types of education. The four types of secondary school are:

- VWO: six-year pre-university schools leading to diplomas entitling pupils to enter University. *Atheneums* concentrate on 'modern' subjects and *Gymnasiums* on classics.
- Havo: Five-year general secondary schools (higher level), whose final exams give pupils access to higher vocational institutions such as teacher training colleges, colleges of education, agricultural and technical institutions.
- Mavo: three or four-year general secondary schools (intermediate level), leading to intermediate vocational institutions.
- LBO: four-year lower vocational schools.



At the end of each of the secondary schools described above, an examination is held.
In how many and which subjects the pupil will be examined is determined during the school career of each individual pupil.
Each of the secondary schools starts with a common programme for all pupils, comprising a large variety of subjects. After several years the pupil has to make a decision about the subjects he will concentrate on, and in which he chooses to be examined at his final examination. The choice between the different subjects is not completely free. The law prescribes for each type of school in how many subjects a pupil will be examined and which subjects are obligatory. For example every pupil has to be examined in his mothertongue and at least one foreign language.
Other restrictions on the choice of the subjects are caused by the demands of the tertiairy educational institutions. Some vocational schools are rather strict as regards the subjects pupils should have in their examination package if they want to present themselves for further vocational schooling. This means that a number of important decisions concerning the future after-secondary school career of a pupil are made one or two years before leaving secondary school.

The assessment procedure

The assessment of pupils for the secondary school certificates is divided into two separate procedures: assessment by means of written, nationally administered examinations and assessment on a local level by the individual school with various means: written papers, tests, project work etc. The results of the two

assessment procedures are averaged for purposes of certification.

The executive responsibility for setting nationally administered examination papers and standards rests with special Committees of Examiners, appointed by the Department of Education. These Committees turn to CITO for help because the Committees do not have the time or the necessary knowledge and experience to develop examination papers for the subject they are charged with that ensure some reliability of judgement. CITO in its turn is commissioned by the Committee of Examiners to write examination papers for the subject in question. CITO then establishes a writing committee, consisting of three experienced teachers. These committees work under the supervision of a CITO-staff-member. Rough drafts of papers have to be submitted to the Committees of Examiners for approval.

CITO also provides the schools with machine scorable answer sheets in those cases that pre-coded answers (Multiple choice etc.) are used. In those cases where open-ended or essay questions are used, the committee charged with the writing of the papers also draws up fairly detailed instructions for marking the papers. In these cases the papers are marked by the pupils' own teacher. This is then checked by a colleague of another school, as the so called 'second marker'. There is no central check on the way in which the markers have used the instructions while doing their work.

In the cases that pre-coded answers are used, CITO scores the answer sheets mechanically and provides the raw scores of the pupils.

Those raw scores are reported to the Committee of Examiners for the purpose of setting standards and establishing grades. In the case of teacher marked papers, data are collected to fulfil the role of raw scores.

The grading is in most cases a compromise between absolute and relative grading. Although certain standards are set by the Committee of Examiners before the results of the pupils are known, the expected percentage of insufficient marks is also taken into consideration.

This picture of the final examinations surely raises certain questions. First there is the big difference between teacher and machine rating. In some subjects we have an objective testing system, for other subjects the teacher is an important factor in grading the pupils. This draws the attention to a certain ambiguity in the whole system: does the teacher have to play an important role in the decision to give a pupil his final certificate or doesn't he? As we have already seen, the image is hybrid, for certain subjects the teacher is an important intermediary, for other subjects he doesn't have any influence in the pupils' marks. We think this ambiguity is made possible by the indistinctness of the function of final examinations. In its recent publication on the policy to be developed in the field of final examinations, the Department of Education formulates no fewer than 8 possible functions examinations can have:

A to select and allocate pupils

B to qualify pupils
C to predict future success in schools
D to guarantee educational standards
E to enable the Government to 'manage' education
F to operationalize educational objectives
G to socialize pupils
H to measure the effect of teaching

It will be clear that the first four functions are relevant to the level of the individual pupil and that the other four functions have more to do with the Government's task to manage the educational system in the most general way. Now we realize that the actual differences between the different examination forms are caused by a different emphasis on one or more of the 8 functions of examinations. It will be clear that an examination that pretends to operationalize all the educational objectives will be different from one that aims mainly at predicting future success in education or society in a reliable way.
CITO's problem in writing examination papers is the fact that the Government had not as yet given clear directions about the functions of examinations that should be focused on. So the decision about the contents of an examination paper and the kind of marking to be used is taken subject by subject, by the Committee in question. There is no question of a guideline on a national level in this matter.

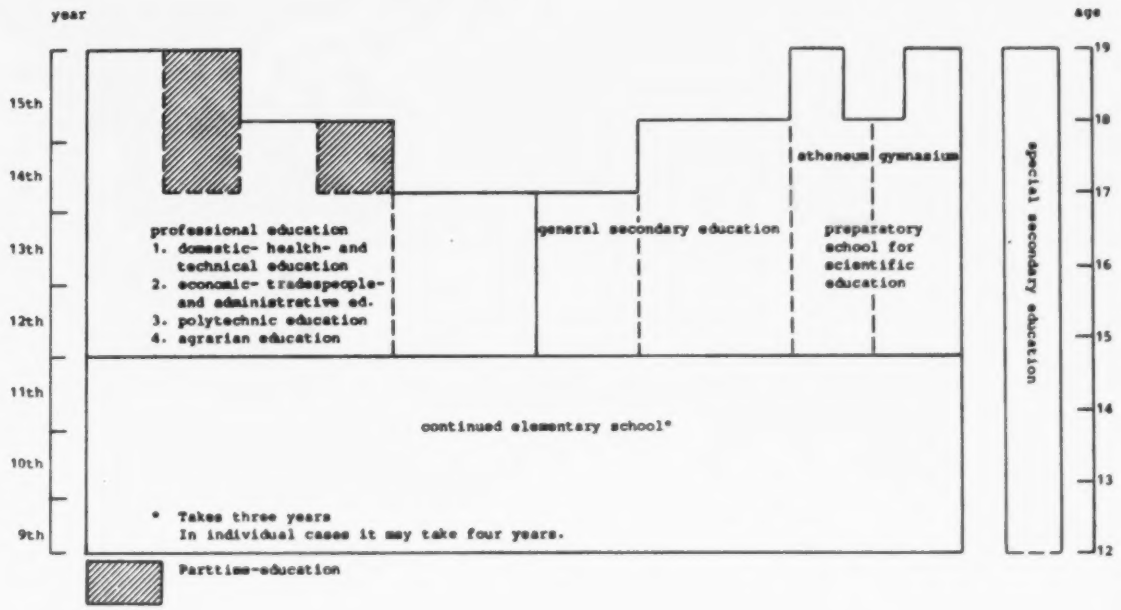
In its recent Examennota though, the Government is proposing some main lines for future development.
In the first place, it is proposed to make a distinction between the school assessment and the central examination. In the central examination the Government wants to stress the management functions, while the school assessment should reflect the teaching that has been done during the schooling period. This would mean that the final examination should measure the centrally relevant part of the objectives while the school assessment should reflect the specific way this particular school has dealt with the objectives. By this measure a teacher could make school assessment tests that closely follow the training in the classroom. So the school assessment could be different from school to school.
In his Examennota the Minister draws the only logical conclusion that follows from this proposal: in the final certification the mark for the school assessment and that for the central examination should be shown separately. No more averageing will take place if these propositions are accepted by Parliament.

Comprehensive secondary school: final examination?

So far I have described the present situation with regard to final examinations and the problem it gives for people involved in the production of the examination papers. As in most other West-European countries systems, we have in the Netherlands at present a rather heated discussion about proposals made by several consecutive governments for a system of comprehensive secondary schools.

This paper is not meant to repeat all the pros and cons the participants in the actual discussions use, so I will limit my contribution to this conference to a general view of the implications the actual Departmental proposals would have for the present certification system.

The structure of the continuation education according to the note: *'Onwards from elementary education'*.



As you see on this schema, the new system would bring us one school for all 12 to 15 year old pupils. The actual secondary schools will only remain in the second phase of the secondary schooling: the 'upper' school. At the end of this 'second phase' of secondary schooling the situation with regard to examinations will probably remain comparable to the actual situation we have described in the beginning of this paper.
But there will be a break in the pupil's career at the age of 15. Pupils have to make a choice between the different possibilities in the second phase.
The present minister has stated that he does not want this break to be marked by centrally administered tests. Two reasons are given of this view: any form of selection, even at the end of the first phase of the secondary school, would be in contradiction of the present minister's view on the essentials of comprehensive secondary schools: the inevitable backwash effect of centrally administered examinations would be contrary to the social-pedagogical ideals of comprehensive schooling. The minister stresses there should be some kind of pupil centered evaluation and counseling but he also stresses the fact that the school itself should take the complete responsibility for this evaluation procedure.
On the other hand the minister doesn't want the Government to neglect its task of caring for the quality of secondary schooling. This brings him to proposing some system of quality control of comprehensive schools, centered on the school as a unity rather than on the learning results of individual pupils.
This strict separation of the different functions that are now all fulfilled by final examinations, implicates a changing view on the function of final examinations. As the proposals of the minister are very recent, I cannot at this moment give you a clear exposé on CITO's view on the development described. We hope to have some time for reflection and study the solutions that have been found in other countries for comparable problems. With regard to this the contribution of Mr Lindblad to this conference is most meaningful to us at this very moment.

Language examinations

Several contributions to this conference are centered on the content of foreign language testing and the testing of the mother-tongue. I will now limit myself to a general view on those examinations, and give you some outlines of developments to be expected in this field.

For any of the languages the pupil chooses to be tested on, the examination is divided into 2 parts: the centrally administered test of reading comprehension and the testing of the 3 other language skills: writing, listening and speaking, a testing that is fully administered by each school itself. The final mark is the average of the school and the central grading. So the central reading comprehension test determines 50% of the pupil's final mark.
The reading comprehension test used for the mother-tongue differs from the ones used for the foreign languages.

Mrs Samson has drawn your attention to the foreign language reading comprehension test, and Mr Kreeft and Mr Sanders will inform you in further detail about the centrally administered testing of the mother-tongue.
I will discuss the more general context of those tests.
As I have told you, the whole system of the division of the testing into a part left to the responsibility of the teachers and a central part is more and more the subject of discussion.
If the trend towards a clear distinction between the functions of the two different examination situations persists, the actual situation with regard to language examining might change. If the school assessment is to have another function than the central examination, there is no reason to persist in reserving certain skills for school-made tests. If the school tests are to reflect the school curriculum and the particular choices a certain school has made within the (nationally fixed) curriculum, the school tests should represent all the past learning activities, without exception for any language skill.
The function of the central examination would then be to check whether the more general objectives prescribed on a national level have been reached. And again, if general objectives are to be described for the 4 skills, the central examination should be representative for all of them. Of course this will place us before the problem of constructing reliable tests for the productive skills too. A problem not so very easy to solve. But it would be the ultimate consequence of the philosophy presented in the minister's Examennota.

So we are faced in this case with possible changes in examinations, resulting from a more general educational philosophy.

But in the case of language we experience other changes these days. You are all familiar with what is called nowadays the notional functional approach in defining language objectives. The actual use of language is getting more and more attention in the setting up of curricula.
Since the Council of Europe has published its first studies on this approach, and I now refer especially to Van Ek's **Threshold Level**, published in 1975*, the influence of the new approach has not stopped to increase. This means that the foreign language curricula as we have known until now will give rise to some critical questions. The one critical remark I would like to draw your attention to at this moment is the question whether we can continue to define foreign language objectives fundamentally in terms of skills. It seems that the 4-skills approach is losing some ground, especially in those situations where it was used in its most strict way. We all realise now that in an actual communication situation especially speaking and listening are closely interdependent. We, as testconstructors, see ourselves faced with the problem to develop new test-procedures that have a higher content validity with regard to the new accents in objectives than our present tests have, and that are on the other hand at least as reliable as the present ones.
Up to this moment the notional-functional approach has

had most of its influence in didactics and the setting up of curricula, and we are only at the beginning of the discussion that will surely follow: the critical review of the content of our present language examination.

I hope I have made clear to you that our examinations are subject to influence from different sides. On the one hand the educational system can never be regarded as a static system, on the contrary, it follows the moves of the society in which it functions. On the other hand the subjects taught in secondary schools are also liable to change.

I hope that meetings like this help us by the confrontation of ideas to find the optimal approach to the solution of the problems outlined in this paper, especially with regard to the form and content of language testing.

---

* Van Ek, Threshold Level in a European unit/credit system for modern language learning by adults. Council of Europe Straatsburg 1975.

Backwash Effects of Multiple-Choice Language Tests: Myth or Reality?

Hildo Wesdorp

1 Introduction

Since the mid 1960's, the use of objective testing in education, in particular of multiple-choice tests, has increased dramatically in the Netherlands. In 1968, the National Institute for Educational Measurement (Cito) was founded and since then the construction of many achievement tests has been concentrated there. This general changeover from traditional testing to objective testing also applies to language teaching. It has resulted in the construction of a fair number of multiple-choice tests, not only for Dutch as a mother-tongue, but also for French, German and English, the three most important foreign languages taught in Dutch schools. Multiple-choice tests for Dutch as a mother-tongue are now part of the final testing program at the end of primary education which plays an important part in the selection for secondary education. Some types of secondary education also incorporate multiple-choice tests for Dutch in the final examinations which determine a student's admission to higher education. The same holds for foreign language tests: they form an important part of the final examinations in various types of secondary education.

From an educational point of view, the introduction of a new type of test for an important selection procedure is certainly not a trivial matter: the new tests are frequently feared for their possible influence on education in general. Most of the time they provoke heated discussions in education journals, but the debate generally subsides after a while, when teachers have become used to the new tests. Eventually, the educational system accepts the new situation and the discussion stops altogether. At first glance, such a development seems perfectly acceptable, but this course of events also has its problematic sides. The developers of the new language tests are primarily interested in perfecting the technical quality of their measuring instrument and do not worry so much about its effects on the educational system. Nevertheless, these effects determine in part the validity of the newly introduced instrument. After all, tests are made to accommodate educational goals and their content-validity is defined as the relationship between the content of the test and of the curriculum. If a test with only moderate content-validity is introduced in a high-status testing situation, for example in an important selection procedure, the chances are that the educational system will adapt to the questions being asked in the test. As time goes on, this process will increase the validity of the test content and, after a while, the fit between the educational system and the test will be perfect, giving the test perfect content-validity, while the original discussion about the discrepancy between

test content and educational goals recedes into the background.
Those who construct tests should be more aware of such developments. Such an awareness should, in the first place, lead to the development of tests with a perfect fit to the existing educational goals, so that backwash effects will be out of the question. In the second place, this awareness should be an incentive for research into the possible effects of introducing the measuring-instrument in question, when there is doubt about the desired fit. In this way, test-constructors would achieve a more balanced view of the various aspects of their testing instruments.

In the heated debate on the effects of the introduction of the new type tests, four objections are most often heard in the Netherlands. First, **discriminatory effects** were expected. The new tests were thought to put certain social groups at a disadvantage. Second, an **impoverishment of the curriculum** was expected. It was thought that skills which could not be measured through multiple-choice tests would not be practiced in the schools and would, therefore, gradually be lost. In the third place an **impoverishment of teaching methodology** was expected. Supposedly, a final language examination with multiple-choice tests would preclude the use of certain teaching methods, causing a gradual decline of their use. And the fourth drawback was the possible **influence of multiple-choice tests on the study-habits of students**. Students were expected to prepare themselves in a different way when used to multiple-choice tests. In addition, certain emotional aspects of multiple-choice tests (e.g. fear of test-taking) were mentioned.

This paper reports on various research-projects carried out during the years 1976-1979. They were all aimed at the question: can we find some of the mentioned backwash-effects? Research on the four groups of effects, mentioned above, will be reported in the sections 2-5.

2 The discriminatory effect of multiple-choice tests (reading comprehension)

From earlier surveys and from the Dutch literature on the subject, it became clear that to many people the introduction of multiple-choice tests in examinations implies an increase in selection on the basis of socio-economic status. Multiple-choice tests are looked upon as undemocratic and are thought to place certain social groups or regions at a disadvantage when such tests are part of an examination.

These opinions often appear to be based on the idea that multiple-choice tests appeal more to certain verbal skills than traditional instruments. This is seen to be a disadvantage for certain social groups, particularly when they are also confronted with middle-class language, of which they do not have a good command.

In this investigation the supposedly discriminatory effect of multiple-choice tests has been assessed in two different ways. One part of the investigation was carried out in the final year of HAVO schools, for English, while the other part was directed at the LBO examinations for Dutch*. The object of the investigation was to find out whether multiple-choice tests discriminate against certain social groups, as compared to traditional tests.

For the first investigation, parallel instruments were constructed which were as much equivalent as possible. One consisted of multiple-choice questions and one of essay questions. Two versions were made of this instrument, so that half of the students were examined first on the multiple-choice questions and then on the essay-questions, while the other half were tested in the opposite order. As far as content was concerned, all students answered the same sixteen questions. The tests, together with questions concerning the occupation of the student's father, were administered to 412 final year students from twelve HAVO-schools.

The average degree of difficulty (p-value) of the multiple-choice questions was .57 and of the open questions .34. The test which first presented the multiple-choice questions, followed by the essay-questions, showed slightly better results than the test presenting the questions in the opposite order. The reliability (Cronbach's α) of both test versions was .63. The average scores for students of different socio-economic status is shown in Table 1. Considering the low reliability of the test, the differences in the scores of Table 1 should be interpreted with caution.

* HAVO: middle stream of secondary education, 5 years.
LBO : lower stream of secondary education (vocational track), 4 years.

Table 1 Mean test-scores HAVO-English

Socio-economic status	N	Multiple-Choice	Open-Q	Difference-scores
1 unskilled and skilled workers	64	4.74	2.40	2.34
2 lower level employees	108	4.43	2.84	1.59
3 small business-owners	45	4.11	2.28	1.83
4 mid-level employees	44	4.68	3.07	1.61
5 high level employees	69	4.46	2.59	1.87

The test results have been analyzed in a number of ways. Separate analyses of variance for both test halves did not show any interaction effect between socio-economic status and question format. An analysis of score differences did not indicate a discriminatory effect for multiple-choice questions, nor did the relative results for boys and girls on essay and multiple-choice questions yield any difference. Significant differences were found between schools, but there is no question of an interaction effect between school and socio-economic status; in other words, none of the schools was found to have a socially discriminating influence.

The results of this first investigation do not in any way indicate an 'anti-democratic effect' for multiple-choice questions. It is, of course, possible that the attempt to assess the effect of the differences in question format as accurately as possible, by means of keeping the question-content constant, has removed the essential differences between the two types of questions from the experiments. The presumed differences between multiple-choice and open questions are not limited to the presence or absence of a list of alternative answers under the question. One criticism often levelled against objective tests is their poorer content as compared to traditional tests.

A comparison of traditional and multiple-choice tests - both representative for their category - was the purpose of the second investigation. The current situation for LBO final exams offered a good opportunity for such a comparison. One track of LBO has a two-part final examination for a number of subjects, including Dutch. The students take one exam, consisting of open questions and one consisting of multiple-choice questions. By collecting both types of results for each candidate, as well as data on socio-economic status, a comparison could be made of the difference in achievement on either type of test for students from various social groups. By using data from actual exams, maximum student motivation and achievement is guaranteed, while artifacts of an experimental situation can also be avoided.

For 258 final-year students from seven LBO-schools the final exam results for Dutch were collected in 1977 and a classification was made of the fathers' occupations. In Table 2 the exam results are given as follows:
- Average scores and standard deviations for the essay-questions on the examinations: maximum score 57.
- Average scores and standard deviations for the

multiple-choice part of the examination: maximum score 50.

Table 2 Mean scores and standard deviations for students of different socio-economic status: final examination Dutch (LTO 177)

		N	Multiple-Choice M	Multiple-Choice SD	Open Questions M	Open Questions SD
1	Unskilled labor	29	26.69	4.67	36.69	6.70
2	Skilled labor	95	28.74	5.27	36.18	6.62
3	Lower level employees	46	28.48	5.51	37.02	5.69
4	Small business-owners	58	29.52	4.85	36.81	6.68
5	Mid-level and high level empl.	30	29.83	5.09	36.43	7.71

The table shows that the students scored higher on the essay questions than on the multiple-choice part of the test, while the standard deviations on the essay questions are also higher. The correlation of .36 between the two formats is significant, but rather low, presumably because both test forms measure a number of skills simultaneously. Neither the open questions nor the multiple-choice questions showed significant differences, according to socio-economic status, although the averages for multiple-choice questions for higher socio-economic status students were somewhat higher. The average differences of the five status groups seem to show the predicted tendency, but there was no question of any significant difference. Since the reliability of the score differences is probably also quite low in this case, given the low reliability of the scores for both question formats, these results should be interpreted with caution. By means of an analysis of variance, a significant difference between the seven schools on both types of questions could be found, but there was no indication of an interaction effect between school and socio-economic status. In other words, none of the schools seemed to show an influence of socio-economic status on the differences in results for the two test formats.

From this investigation it may be concluded that a socially discriminating effect for multiple-choice tests in the 1977 final examination for Dutch in LBO-schools and in the constructed HAVO-tests for English could not be demonstrated. Achievements on multiple-choice and open question-ended tests for Dutch and English were not found to be different for the various participating social groups.

In a survey among foreign language teachers about their attitudes towards multiple-choice tests we found that considerable minorities of teachers believe in the discriminatory effects of the multiple-choice reading comprehension tests, that are part of the final examinations. A statement like 'Reading comprehension tests (in multiple-choice format) are undemocratic; certain social classes or regional groups are put at a disadvantage' was supported by 20% of the teachers. And a statement like 'Because of the introduction of multiple-choice reading comprehension tests in the examinations the selection according to social class has been intensified' was supported by 18% of the teachers.

Although our investigations did not find any proof of such socially discriminating effects, important minorities of teachers still strongly believe the opposite. Research findings and teacher opinions are conflicting.

3.1 The influence of multiple-choice testing on the writing skills of students

In 1966, the first achievement test was administered in Amsterdam in the final year of primary education. This achievement test has slowly acquired a certain degree of status, because its results determine in part in which type of secondary education the student will be placed. The instrument consists entirely of multiple-choice tests on all the important subjects taught in primary school. For Dutch as a mother tongue the test contains subtests on reading comprehension, spelling, grammar, and an attempt to measure writing skills in an indirect way. An essay containing various mistakes is used to assess a number of writing skills. The student is requested to correct the mistakes, choosing one from a few alternatives. The mistakes concern skills in a variety of areas. These areas are: **writing conventions**, i.e. dealing with problems of spelling and punctuation; **syntax, appropriate and effective language-use**, i.e. stylistic nuances; **organization of the essay**, i.e. dealing with structural mistakes, such as lack of necessary information, poor use of paragraphs; and **content**, i.e. mistakes in the content of the essay. The first year in which the exam was given (1966), all participants also wrote an essay, which was not evaluated in any way.

It was thought that the limited scope of the test content (only multiple-choice items, no essay questions, no testing of productive language skills) would lead to less attention being paid to independent essay-writing, eventually leading to a decrease in writing skills on the part of the students. The 1966 requirement to write an additional essay made it possible to compare writing skills in 1978 and 1966. A comparable sample of students was required to write an essay, similar to the one in 1966. The two sets of essays were then compared for quality, presupposing that the effect of twelve years of multiple-choice testing would have resulted in poorer quality of the 1978 essays.

Two comparisons were made in this investigation. First, 64 'old' and 64 'new' essays were compared for overall quality by 15 judges, giving a global judgement. Obviously, there were no differences in format between the old and the new essays and they were mixed randomly. Secondly, 32 old and 32 new essays were judged analytically: i.e. judgements on 'content', 'style', and 'writing conventions' were elicited from 16 (different) judges. Differences were expected, both in the comparison of the global judgements and of the analytic judgements, which could presumably be explained by the 'impoverishment-effect' of the achievement test. The

reliability of the jury-judgements (the degree of consensus with other juries of equal size and competence) varied from .78 to .95 for the various criteria. Table 3 shows some results of the investigation. The global judgement the total analytic judgements were given on a ten-point scale, the type of scale generally used in schools in the Netherlands, while the analytic categories were judged on a four-point scale.

Table 3 Mean scores and standard deviations for old (1966) and new (1977) essays, by means of global and analytic judgement by juries of 15 and 16 members, respectively

	GLOBAL		ANALYTIC							
			Total		Content		Style		Conventions	
	Old	New	Old	New	Old	New	Old	New	Old	New
M	6.50	6.60	5.96	6.04	2.64	2.77	2.84	2.79	2.95	2.80
SD	0.68	0.67	0.67	1.02	0.38	0.56	0.44	0.57	0.52	0.61
n	64	64	32	32	32	32	32	32	32	32

None of the expected differences was actually found in the investigation, leading to the conclusion that no impoverishment as a result of the use of multiple-choice achievement tests is to be found in writing skills of the pupils. Enrichment in the more technical aspects of writing skills, such as 'writing conventions' could not be found either, even though the test does especially stress such points every year.

3.2 A comparison of old and new reading comprehension exams for the mother-tongue at the end of secondary education

For one type of secondary education, Lower Technical Schools (LTO), it was possible to compare traditional reading comprehension exams using a text with open questions, with the new form of reading comprehension tests using a text followed by a number of multiple-choice questions. For a number of years, these two examination formats existed side by side, which made it possible to compare the skills that were being measured by the two types of tests between 1971 and 1977. This was done by using the criteria for measuring reading comprehension skills as formulated by F.B. Davis in: The literature of research in reading with emphasis on models. Rutgers; The State University, New Brunswick (N.J.), 1971

Nine operational skills of comprehension in reading among mature readers

1 Word knowledge, as measured by recognition vocabulary items
2 Ability to select the appropriate meaning for a word or phrase in the light of its particular contextual setting

3 Ability to follow the organization of a passage and to identify antecendents and find references in the passage
4 Ability to select the main thought of a passage
5 Ability to answer questions which are answered directly in the passage
6 Ability to answer questions which are answered in the passage but not in the words in which the question is asked
7 Ability to draw inferences from the passage about the content of the passage
8 Ability to recognize the literary devices used in a passage and to apprehend its tone and mood
9 Ability to determine the writer's purpose, intent and point of view; i.e., to draw inferences about the author

Table 4 gives the percentage for each category of the Davis system which is required for answering the two types of exam question, according to the judgements of a three-member jury. The open question exams contained a total of 276 questions, the multiple-choice exams contained 447 questions. The average inter-judge reliability was 57%, which is why the results for the three judges have not been collapsed into one figure. According to all three judges, there are differences in the way the skills for the two types of examinations are divided among the Davis categories.

Table 4 Classification of traditional and objective examinations for Dutch reading comprehension for LTO-schools from 1971 to 1977, by three judges, according to the Davis criteria (in percentages)

Examination	Categories									Not
Open Question	1	2	3	4	5	6	7	8	9	classified
judge I	20	7	20	3	19	17	8	6	0	0
judge II	12	13	16	0	31	10	6	2	0	9
judge III	24	11	19	0	32	6	5	0	0	4
Multiple-choice										
judge I	6	11	6	10	23	26	9	6	3	0
judge II	1	18	7	3	31	16	18	4	3	0
judge III	4	16	10	4	27	20	13	4	2	1

The three judges found that on open question exams, questions about word knowledge (Davis skill 1), and grammatical structure and function (skill 3) are asked more frequently. Both multiple-choice and open questions exams frequently demand answers that are found directly in the passage (skill 5). Multiple-choice questions often require the ability to select appropriate meaning from context, recognize parahrases, or draw inferences (skills, 2, 4, 6, 7 and 8). In order to see to what extent these differences are due to a general change in exams over the period from 1971 to 1977, it was decided to compare the Davis classification for the last three

exams separately (two exams from 1976 and one from 1977). Table 5 shows the required skills in percentages, according to the judgements from three judges.

Table 5 Classification of traditional and objective examinations for Dutch reading comprehension for 1976 and 1977 by three judges, according to the Davis criteria (in percentages)

Examination	Categories								
Open Questions	1	2	3	4	5	6	7	8	9
judge I	1	3	10	5	41	27	10	2	1
judge II	0	2	13	0	62	15	7	1	1
judge III	0	10	13	0	65	10	1	0	0
Multiple-choice									
judge I	4	9	3	11	25	26	8	9	5
judge II	0	16	5	5	30	21	13	4	5
judge III	4	13	7	6	29	20	11	5	4

Testing the frequency-distribution of the various skills for the two exam types yielded significant values for all three judges. According to all three judges, skills 6, 7, 8 and 9 were required more in multiple-choice tests. This means that, according to the judges, the exams in the last two years required more ability to paraphrase, draw inferences, recognize literary devices, and determine the writer's intent and point of view. In 1976 and 1977, both multiple-choice and open-question exams required less word knowledge than did earlier exams.

The conclusion that may be drawn from this investigation is that the analyzed LTO-reading comprehension examinations have not become poorer in content as a result of the introduction of multiple-choice tests. On the contrary: a certain degree of enrichment can be found in comparison with the open question examinations. Inasmuch as the differences between the two types of exams have become smaller in recent years, or have disappeared completely, this seems to have resulted from an improvement in the essay questions after the introduction of multiple-choice questions in final exams.

In a survey among foreign language teachers we also asked for their opinions on the subject of the impoverishment of the curriculum, especially the foreign language curriculum in the various streams of our Dutch secondary system. The teachers turned out to be rather critical. Important minorities supported the following statements:
The foreign language reading comprehension tests put too much stress on general reasoning questions, whilst the knowledge of the second language is of secondary importance (67%).
The foreign language reading comprehension test neglect:
- lyrical texts, emotion-evoking textelements (66%)
- reacting to texts, criticizing the texts, text analysis, evaluating the texts (81%)
- humorous, ironical, witty texts (48%).

Our two empirical investigations were both aimed at the discovery of effects on the mother-tongue curriculum. We could not find any evidence of the existence of such effects. Our foreign language teachers however, are quite sure that in their field the final examinations in multiple-choice format lack validity and therefore must have certain backwash effects.

4.1 The influence of 11+ tests on mother-tongue teaching in the primary schools

Another investigation was designed to see whether the achievement test mentioned in 3.1, with its limited coverage of the goals of mother-tongue education, had any effect on teaching practices. To this end, a comparison was made of the time spent on various parts of the curriculum and of types of teaching in 102 schools. 51 schools had been using the test for quite some time and 51 had not used the test at all. Obviously, the expected differences were that the 'test schools' would pay more attention to components of the curriculum that are part of the test and less to components that will not be tested, in comparison with the 'non-test schools'. Two types of information were elicited by means of a questionnaire: an assessment of the time spent on various types of teaching and of the teacher's attention to different parts of the language curriculum.

Table 6 Time spent on various teaching methods in 51 'test-schools' and in 51 non-test schools' a)

type of teaching	test schools n	test schools med.	non-test schools n	non-test schools med
teaching in class	47	3.86	50	3.25
teaching individually	48	2.07	50	2.17
teaching small groups	47	2.00	50	1.62
teacher activities				
teaching	44	2.21	41	0.88
narrating	50	1.63	48	1.29
explaining	51	1.74	48	1.90
demonstrating	49	0.98	45	1.06
individual activity	50	2.17	47	2.00
class discussion	51	1.07	49	1.16
oral quizzing	48	0.95	46	0.67
written quizzing	50	1.07	48	0.90

a) Time spent is expressed on a seven-point scale, ranging from 1 (almost never) to 7 (more than twenty hours per week).

Table 6 shows that few differences could be found. The differences were subjected to the Mann-Whitney U-test, which showed that both oral and written quizzing is done more frequently in the 'test schools'. Table 7 shows the way in which the teacher's attention was divided among the various parts of the language curriculum.

Table 7 Time spent on language curriculum components in 51 test-schools and non-test schools b)

Curriculum component	Part of test?	Test schools n	Test schools med.	Non-test schools n	Non-test schools med.
Technical reading	no	51	1.00	50	1.07
Reading comprehension	yes	51	1.36	50	1.23
Essay writing	no	51	0.86	49	0.84
Grammar	optional	48	0.97	48	1.01
Spelling	yes	51	1.25	49	1.13
Style excercises	yes	49	1.14	49	1.16
Listening comprehension	no	49	0.97	48	0.97
Speaking	no	50	0.95	50	1.00

No differences could be traced in the time spent by teachers on various parts of the curriculum, even though some differences had been expected. Consequently, the investigation led to the conclusion that the use of achievement tests has no clear influence on teaching practices.

4.2 The use of multiple-choice tests (as teaching aids) in secondary education

The degree of test-use in the classroom could be considered as an indicator of possible influence of multiple-choice final examinations on the curriculum. The more frequently multiple-choice tests are used in class, the more likely it is that teaching-content is reduced to teaching only what will be tested in the final examination. In this section a summary is given of a survey of the frequency with which multiple-choice tests are used for teaching in the final years of secondary education, as well as the other uses to which such tests are put.

A questionnaire was sent a to sample of schools for secondary education, in particular to the departments whose subjects are partially or completely tested by means of a multiple-choice test in the written final examination. 965 schools were approached, comprising 4471 departments. The percentage of responses was 47% (2102 questionnaires). The survey yielded a large amount of interesting data, which cannot all be summarized here, but the most striking results for language teaching will be given below.

The first part of the survey was concerned with the format of test questions used in the classroom. In the final year, the percentage of multiple-choice questions compared to essay and short-answer questions was 21% and 15% respectively for French, 23% and 11% for German, and 27% and 14% for English. For Dutch the percentages were higher: 34% compared to 30%. These figures imply that multiple-choice tests only have limited influence on classroom test format, while the influence is greatest for Dutch.

The second part of the survey was designed to find out

whether multiple-choice tests are used as teaching aids in the classroom and as homework assignments. The results for language teaching have been summarized in Table 8 and show that multiple-choice tests are not being used on a large scale. Special exercises to train students for multiple-choice questions are given frequently, though.

Table 8 Use of multiple-choice questions in the classroom and as homework-assignment in the two final years of secondary school, in percentages

	Dutch	French	German	English
Multiple-choice questions used as teaching aid in teaching subject material	24	20	19	25
Students given multiple-choice questions in class for practising subject material	52	34	31	38
Students given multiple-choice questions in class for practising multiple-choice questions	58	40	54	51
Students given multiple-choice questions as homework for practising subject material	27	24	21	27
Students given multiple-choice questions as homework for practising multiple-choice questions	31	22	29	29

In the same survey among foreign language teachers mentioned before we also asked teachers about their opinions on the influences of multiple-choice testing on foreign language teaching methodology. Most teachers are rather critical: majorities support statements like the following:

- Reading comprehension tests occupy too important a position as didactic aids (68%).
- I am speaking much more Dutch now than the language I teach (54%).
- Testing the various skills separately leads to teaching them separately (58%).
- Because of the final examinations in multiple-choice format teachers are forced to use a traditional 'frontal' class-teaching approach (59%).

But on the other hand foreign language teachers seem not so very impressed by the impact of the reading comprehension tests: 62% still think that there is only a slight chance that teaching methodology will become rigid and uniform.

Nevertheless, we still find the same phenomenon: while our empirical research cannot really prove that there are effects on teaching methodology, the teachers are quite sure they exist. The conflict between research findings and teacher opinions arises again.

5 Influences on the study habits of the pupils

Many kinds of influence on education from tests and examinations have been studied, but the influence of tests on student's study methods has been studied more than most. In a large body of investigations students have answered questionnaires stating that they study differently for multiple-choice tests than they do for traditional - open question - exams. Most of these investigations have been directed at factual school-subjects. An earlier Dutch survey among foreign language teachers showed that 43% of the teachers perceive student's study habits for foreign languages as being influenced unfavourably by the use of multiple-choice reading comprehension tests, both in exams and in the classroom.

In order to study this problem, an inventory was made in 10 Havo final-year classes of the study methods used by students who are either frequently confronted with multiple-choice tests, or with traditional testing instruments. The teachers in the participating schools had earlier been asked, how many times per month and per year they gave their students multiple-choice tests and how often open question tests. The frequencies reported by the teachers were used to form two groups of students for each subject, one with relatively little and one with relatively substantial experience with multiple-choice tests. After taking the tests, students were given a questionnaire about their preparation methods for open question or multiple-choice tests for English, depending on their reported experience with multiple-choice tests. Students who did not do any preparation at all did not fill out the questionnaire. 99 students from 5 schools were given a questionnaire about multiple-choice tests, while 85 students from another 5 schools filled out a questionnaire about preparing for open-question tests. The questionnaires contained 11 questions about preparation methods for tests, one question about student preference, and one question about the type of test-questions students found most difficult.

The experience students claimed they had had with open question tests did not always coincide with the frequency reported by the teachers. To a lesser extent, this was also the case with multiple-choice tests. Most students report little experience with open question tests: only 17 % report more than one such test per year. Students appear to do little or no preparatory work for English tests. For a multiple-choice test, 2% of the students study 'always' or 'most of the time', 20% 'sometimes', and 78% '(almost) never'. For an open question test, these percentages are 32%, 20% and 48%, respectively. 84% of the students spend an average of less than half an hour preparing for a multiple-choice test, while for an open-question test 46% prepare less than half an hour, 29% between half an hour

and an hour, and 24% more than an hour. The preparation time for open question tests is mostly spent on learning words and expressions and looking through lists of idioms. Old reading comprehension exercises and tests are less popular as study material. The answers showed a strong student preference for multiple-choice tests (72%).

Although students still study a little longer for open-question tests, they generally do not prepare much at all for English reading comprehension tests. Within the group with more multiple-choice experience, a strong correlation was found between the frequency of being tested on multiple-choice tests and the degree of preparation.

In a questionnaire presented to samples of approximately 200 students for the subjects Dutch, English, and German, we also asked some questions related to the subject of this section. The following seems to be of importance:

Multiple-choice tests and the student's self-image. Is taking multiple-choice tests the best way for students to gain insight into their own capabilities? Students do not express an opinion about this aspect. About half of the students recognize the positive feedback qualities of multiple-choice tests (for Dutch and English). For German they are more negative on this score: only a minority believe that taking multiple-choice tests tells them anything about their own level of knowledge.

Multiple-choice tests and fear of test-taking. A certain degree of test-taking anxiety is certainly produced by multiple-choice tests: 70% of the students disagree with the statement that multiple-choice tests 'let you know where you stand'. In general, they find multiple-choice tests full of treacherous pitfalls, but only a minority find multiple-choice questions so confusing as to lower the student's performance.

In our survey among foreign language teachers (mentioned above), we also asked them about their opinions on the effects of multiple-choice testing on the behaviour of the students. Large minorities were rather critical: they supported the following statements:

- The thinking pattern of the students will be unfavourably affected (39%).
- Reading comprehension tests encourage a 'fragmented' style of reading, and require too much attention to detail (50%).
- The study behaviour of pupils is disadvantageously influenced (43%).
- We teach out students to become superficial, shallow readers (19%).

Here we also find a conflict between teacher opinions and other research findings. If we believe what teachers think and say, the multiple-choice tests have rather harmful effects, especially on the ultimate goals of education: the way pupils read a text. Our other research however does not support that opinion.

On the basis of the research reported here, the overall conclusion may be drawn that in the Netherlands many of the initial fears about multiple-choice testing have turned out not to be based on facts. Although the fears among teachers more or less remain, we can only conclude that the so-called backwash effects are a myth. If they do exist, they must be so weak or small that our research methods cannot detect them. The Dutch system may have struck the proper balance between multiple-choice and traditional testing, by continuing the use of other testing techniques side by side with multiple-choice tests. This has minimized the risk of educational 'impoverishment'. It is still necessary, though, to carefully follow the developments in this area, in order to prevent an adaptation of the educational system to the testing technique.

At the Dutch national examinations reading comprehension tests are used for all languages. As regards the mother tongue, reading comprehension is tested by means of open-ended questions. Apart from the host of problems that arise at constructing the tests, such as:
- minding its context validity
- determining its degree of difficulty
- avoiding ambiguity of formulation
we also face the problem of rater reliability, which necessitates development of model responses and rating schemes. We will discuss some research we have done these last years on the rater reliability to be achieved with the help of different (i.e. more or less detailed) model responses. Experiments have been carried out and the results of the actual examinations have been analysed. One of the main conclusions is that the influence of the raters' own standards were not controlled by the model responses provided. No matter how detailed the modelresponse used, the rater reliability of the open-ended questions exams was always rather low. Further research on how to control the influence of the raters' own standards is needed. Selecting and training raters for the national examinations might be considered.

In the Dutch national examinations two types of tests are used for the various subjects taught at school: the multiple-choice test and the test with open-ended questions. The reasons for this are based on political and traditional rather than on domainreferenced or testing criteria.

The multiple-choice exams are scored mechanically. Open-ended exams are scored by the students' own teacher and a second rater, a teacher at another school, who has been appointed by the authorities.
Compared to, for instance, England, this situation is rather unique: in the case of national examinations using open-ended questions the work of the candidates is never rated by their own teachers, but by officially appointed raters.

In recent years Cito has done some research on the problems connected with open-ended exams. In this introduction I should like to discuss three studies concerning open-ended reading tests in the mother tongue, in which interrater reliability plays a part.
From what follows it will become apparent that all three studies concentrate on test reliability and the role of raters in this respect. But the results as such cannot be compared because the research questions and analytical techniques differed considerably. A reason for presenting them together in this context is the fact that they give a fairly clear picture of the subject.

The first study concentrates on the aptitude of various types of rating-models for an experimental open-ended reading comprehension test.
The second study which is still in progress tries to develop an analytical rating model for a reading comprehension test comprising only one assignment: Give an approximately 400 word précis (summary) of a text of about 2500 words.
The third study focuses on test-reliability, including interrater reliability, as found in an actual open-ended reading comprehension exam.

A comparison will be made with similar studies on subjects such as physics, economics, music and geography. It is not possible to make a comparison with reading comprehension tests in foreign languages, such as French, German and English, as these tests are multiple-choice tests so that interrater reliability does not apply.

Question types

Before presenting the three individual studies, I think it is necessary to tell you something about the types of questions and rating-models we use for the various exams.

When considering the 'open'-ended questions that are used in the reading comprehension examinations we can see that the term 'open' perhaps is not correct for most of the questions. As a matter of fact many questions are rather closed. See for example questions 1, 2, 3 en 4. The answers we expect are often fully predictable. Many questions can be answered by giving a quotation from the text (question 1 and 2). Even when the question requires an answer to be formulated by the candidate this answer can usually be predicted by the testconstructor. So, when I speak of open-ended questions, I only mean that the candidate must write down the answer himself, unlike true-false questions, multiple-choice questions and matching questions. Real open-ended questions, which demand a non-predictable response from the candidate are hardly ever used.

In the appendix I have included some examples of questions as used in the examination of reading comprehension in the mother tongue. The text I used is an original English text from a second language examination. In the real examinations for the mothertongue the texts are much longer and the number of questions of course greater.

The first two questions are short-answer questions which are to be answered by quotation from the text or by indicating the passage in which the problem mentioned in the question is discussed.
The third question must be answered in the candidates' own words because it concerns an element of the text which is not explicit.
The fourth question is centered on the main idea of a passage in the text and should also be answered in the candidates' own words.

The fifth question asks for the candidate's judgement on a passage in the text (intra textual criticism) the question can be answered in different ways. The criterion here is not whether the answer is right or wrong, but the consistency between judgement and justification as well as the completeness of the answer.
The sixth question shows what a précis (summary) assignment is like. It is used for the testing of reading comprehension at pre-university level. Here too the original texts are much longer and so are the rating models (grids). In this assignment the following elements are tested:

a Has the student included the major points in the text and the relationships between them? (content factor)
b Is the students' phrasing correct? (the phrasing factor)
c Has the student structured his précis in a correct and well-ordered way? (the structure factor)

When testing with open-ended questions we have to face the problem of rater-reliability or rather unreliability. This means that our attention must also be directed to model responses and rating schemes. The position we take is that an open-ended question not finished after the question has been formulated. An open-ended question consists of the question and a detailed model response or rating-scheme.
A model response does not only consist of answers to be judged as right, but also of answers that are partly right, answers that are wrong and instructions about the deduction of scores.
A rating-scheme includes criteria according to which an assessment is made. And a description of the character of the answers going with each score.
Every type of question requires a specially adapted model for the assessment: multiple-choice questions and some types of short-answer questions and only need a correction key. Other short-answer questions and more extended questions require either a model response which lists all possible answers explicitly or a model response which gives indications as to the correctness of the answer in a more abstract form. For all those types of questions I have given examples in the appendix.

<u>The first study</u>: interrater reliability with a reading test in the mothertongue with open-ended questions at O-level.
In 1979 we did some research on rater reliability obtained by using various questions and model responses. Literature on rater reliability indicates that the rater reliability for open-ended questions is generally rather low. The accumulated evidence leads to three inevitable conclusions:

a different raters tend to assign different grades to the same paper
b a single rater tends to assign different grades to the same paper on different occasions
c the differences tend to increase as the essay questions permit greater freedom of response.

Most of the research however concerns the essay type

question. In the Dutch reading comprehension tests however we frequently use a question type in between short-answer questions and essay questions, which we call extended-answer questions. We wanted to know if the results of the research to be found in the literature concerned, also applied to these extended-answer questions.
Thus we constructed a test with different types of questions and different types of model responses:
short-answer questions - with explicit model responses
extended-answer questions - with explicit model responses
and extended-answer questions with more abstract model responses.

Of each reponse model we constructed two forms.
model response I with correct answers only and one or two answers which are partly correct. This model response was devised by the constructors of the test without knowing the actual answers of the candidates.
model response II which was more detailed and constructed after studying the actual answers of sixty candidates obtained by pretest.
We had two hypotheses:
1 the interrater reliability would be higher using the more detailed model responses than in the case of using the more simple model responses
2 the interrater reliability would be different for each type of question with both reponse models.

I shall not present the whole research plan here. I will restrict myself to saying that we worked with an experimental group of eight raters, thirty answers of candidates to each question. The experimental group judged the answers with model response I and three months later with model response II. A control group of 8 raters judged the answers with the more detailed reponse model II only. Interrater reliability was defined as a ratio of components of variance. The numerator of the ratio consists of the person-component. The denominator consists of the person-component the rater-component and the error-component.

These were our conclusions:
- There was no difference between the interrater reliability in using model response I or II
- Short-answer questions that only requested a quotation from the text had a rather high rater reliability .75 or more
- All questions that concern passages in the text of one sentence, a group of words or one word - irrespective of the type of question - had a rather high rater reliability, .75 or more
- All questions that concern more extended text passages had rather a low reliability of about .50
- It was interesting to learn that there was no difference in the interrater reliability between questions that had to be answered in the candidates' own words no matter if they were really open - (the criterion can't be whether the answer is right or wrong) - or if the answers were predictable
- Analysis of some specific questions showed that deduction rules (see appendix) phrased in general terms

had a negative effect on the interrater reliability and that deduction rules phrased in more explicit terms had a positive influence
- One of the most important conclusions was that the raters' own standards were not controlled by the model responses. The data showed raters who were rather consistently generous and raters who were rather consistently strict.

The fact that the results of the more detailed model response II, did not differ from the more simple model response I could be explained in many ways. Some questions for example could have been formulated in a better way. Some model responses became too extended and complex, and last but not least we did not know anything about the motivation of the raters. But in spite of all these problems I think it cannot be denied that making the model responses more detailed without any additional measures is not effective. The National examination could be the recommendation of tests with closed-questions (multiple choice, short-answer questions).
Other considerations however make this kind of decision inacceptable. We think of the validity of the test, the backwash effects of the test in the classroom and the problems of acceptation by the classroom teachers.
We suggest other possibilities:
- the questions, including extended-answer questions and essay questions, could be more structured, for example by adding response restrictions
- further research on the degree in which the raters own standards can be influenced. We would then know which instructions for raters would be more succesful than the instructions we already use.
- the training of raters could also be considered.

Similar studies were made with regard to physics and economics. In the case of physics some encouraging results were found. Assessment without the use of a model response on the basis of a short list of scores per question resulted in an interrater reliability of .70 and higher. The assessment on the basis of a more detailed model response resulted in an even higher interrater reliability.
In the case of economics however results similar to those in the Dutch reading comprehension test were found. There was no difference in assessment with the more detailed and the more simple reponse model.
In the case of economics half the questions had an interrater reliability of .70 or higher, one quater of the questions had an interrater reliability between .50 and .70 and another quarter even below .50.
These differences can only be explained by the differences between the types of questions the subjects give rise to.

In a second study of a model response the assessment of préciswriting was looked at. In an A-level préciswriting test a text of about 2500 words has to be summarized in 400 words. In the assessment of the results a schematic form of the text is used (see appendix, question 6).
For research purposes a model response was divised which consisted of a <u>text-specific part</u>: statements taken from

the text and their mutual relationships and two general parts in which phrasing and structure can be assessed on 5-point scales.
The rater reliability of 5 different raters was calculated on the basis of the mean intercorrelation of the raterscores lengthened by using the Spearman Brown formula.
The rater reliability obtained by using the text-specific analytical part of the model response was high: .80 which came close to the reliability in the case of global or jury assessment.
There was less agreement on the interpretation of the general criteria with regard to phrasing and structure. The rater reliability on these criteria was very low: between .20 and .40.
In a follow up study an attempt it will be tried to relate these criteria to the text specific model response.
It is true that the form of the exam question differs considerably from the exam questions in the studies referred to above. And the same holds good for the ways in which the coëfficients of the interrater reliability have been calculated. But the conclusion is the same: an acceptable rater reliability is out of the question when parts of the rating schemes ask for personal interpretation by the rater.
The experiences from the studies I have just described, were incorporated in the examinations and model responses. In a third study which I want to discuss now, the effect of the unreliability caused by the factor rater as one of the factors which influence the test-reliability of an open-ended test - the other factor being the candidates and questions - was studied.

In this study an O-level reading comprehension test of the mothertongue, that had been part of the 1980 national examination was looked at. The model response of the examination can be characterized as detailed. The work of 300 candidates was assessed by 15 pairs of independent raters, who each assessed the work of 20 candidates.

On the scores obtained in this way, an analysis of variance was carried out in which the candidates, the questions and the raters were looked at. The part of the observed variance which was caused by the factor rater appeared to be small, compared to the effect of the other factors, particularly the interaction between candidates and questions. The most important conclusion therefore is that the lack of test reliability is caused first of all by the other factors, rather than by disagreement of the raters. The reliability of the examaverages around .60; when the factor rater is incorporated the reliability averages .54. The judgement by one rater instead of two raters results in a test reliability of around .47.
Trying to enhance the reliability of these reading comprehension exams by resorting to model responses and multiple ratings is therefore hardly worthwile. A similar conclusion became even more manifest in a study of the assessment of music and geography (Doevendans, Post, Sanders, 1982). This study did not concentrate on rater reliability or test reliability but on comparison with a criterion. Scores on separate questions were not used in

the analysis: the emphasis was on total scores on the exams as a whole. Three assessment contexts were compared to a criterion, in this case: assessment by the classroomteacher and a second rater who knows the first rater's results, assessment by the classroomteacher only and an assessment in which an average was taken of the teacher's scores and the score of an independent rater. The average judgement of five independent raters was used as a criterion. The surprising result was that the total scores of the three assessment contexts hardly deviated from the criterion, even the judgement by just the classroomteacher. The results could be explained by the fact that the classroomteacher knew that a second rating would take place, the fact that detailed model responses were used and the fact that the exams mainly consisted of short-answer questions, which allowed only little variation between the answers.

When the results of the studies discussed above are considered, we arrive at a number of conclusions, which are hardly surprising:
the degree of interrater reliability depends on the questiontype which is determined by the subject.
But we also arrive at conclusions which demand further research or which could well be applied in our examinations.

- Parts of model responses specifically geared to the work which has to be assessed, have a positive effect on interrater reliability, while parts of model responses phrased in general terms have a contrary effect.

But this is small comfort to those who want to enhance the reliability of mother tongue examinations.

- We have seen that the lack of test reliability was not caused first of all by the disagreement of the raters, but that an extension of the number of questions will probably have a greater effect.

The fact that, in spite of model responses, raters stick to their own standards, that in every group of raters there are both mild and strict raters, may have hardly any effect on the examination as a whole, to the individual candidate it can be of the greatest importance. So further research on possibilities to influence these rater characteristics is needed.

I Television tapes, we believe, will give posterity a much better picture of 1974 than we have of, say, the 17th century, despite those indefatigable diarists. Yet the box does distort the image of the age in some ways.

II Size is the simplest illustration. When it all started the criticism heard most frequently was that the screen was too small and that the public, accustomed to the giants of the cinema screen, would never accept this world of dwarfs. In the early stages it was vain for insiders to argue that everything depended on distance from the screen and the angle of vision.

III Gradually the converted ceased to worry about the smallness but it remains a factor in producing curious effects.

IV The zoom lens and magnification produce other odd deceptions. Football pitches and, more particularly, cricket fields, tend to be foreshortened. People in the pavilion seem on top of the pitch. Creepy-crawlies become nightmare monsters, and the population revealed in a drop of pure drinking-water is a terrific boost for booze.

V These things are unimportant, it could be urged, and future generations will make allowances since the camera cannot lie. But can't it? It can't say anything at all if it is not there, a point that has struck me forcibly when watching the "World at War" series. I was impressed when the commentary honestly acknowledged that the scene as the two arms of the Russian pincers met round Stalingrad had been re-enacted but it made me wonder still more about the authenticity of other incidents.

VI When I discussed the question of screen distortion with a television engineer, he surprised me by saying that the spoken word was probably more authentic than the vision. This, he thought, was because producers were constantly seeking artistic effects and experimenting with angles that did not give a true perspective.

L. Marsland Gander in the Daily Telegraph

1 Quote a sentence from paragraphs I or II of which "Size" (line 4) is an illustration.

2 In paragraph III the writer mentions "curious effects".
In which other paragraph or paragraphs does the writer mention factors accounting for "curious effects"?

3 Explain that what the writer describes in paragraph IV is not a "lie" as meant in line 16.
Don't use more than 20 words.

4 What is the main idea of the passage?
Don't use more than 15 words.

5 a How do you judge the necessity of the second paragraph for the passage as a whole?

b State your reasons for your judgement.
Don't use more than 40 words.

6 Summarize this text in 100 words at most.

Model response

question 1

Yet the box does distort the image of the age in some ways	10 points

no points:

any other sentence

question 2

- *paragraph II*	5 points
- *paragraph IV*	5 points

deduction:

deduct 4 points if more than two paragraphs are mentioned

question 3

the essence of a right answer should be:

the author speaks of a lie when you don't know whether what is shown on television is reality or not 10 points

for example:

partly correct:

only a re-enacted incident, that is not presented as a simulation, is what the writer calls a lie 5 points

only a re-enacted incident is what the writer calls a lie 2 points

deduction:

deduct 2 points if more than 25 words are used

no points:

- *on television many tricks are used*
- *the use of a zoomlens*
- *enlargements*
- *the size of the screen*

question 4

The essence of a correct answer should be:

Television distorts reality 10 points

for example:

- *Television doesn't give a full authentic image of the incidents* 10 points
- *Our opinion that television gives an authentic image of reality is not right* 10 points
- *Television distorts the image of the age in some ways* 10 points

---

partly correct:

- *Television doesn't give a much better picture of the incidents that occur nowadays than diarists did of the 17th century* 7 points
- *With television you can never tell whether you are watching a re-enacted scene (of simulated incidents) or not* 5 points
- *Television uses many tricks* 3 points

deduction:

deduct 2 points if more than 20 words are used.

question 5

I a) *The second paragraph is needless*
b) A complete reason for this judgement should consist of:

1 *the main idea of the passage:*
2 *a comment on:*

the size of the screen or people worried about the smallness of the screen in the prime years of television

The relation between these two elements can be indicated as: when the main idea is 1. than 2. is unimportant.

for example:

Because the author wants to tell us that nowadays television can show reality in a special perspective. In that case it is unimportant to know that in the prime years of television people worried about the smallness of the screen. 10 points

---

II a) *The second paragraph is very useful*
b) The essence of a complete reason for this judgement should be:

The author first presents a line of arguments he doesn't endorse, in this line of arguments paragrah II has a function. 10 points

---

deduction:

deduct 2 points if more than 50 words are used.

Question 6

line	Elements of information, structure and relationships*
1	Thesis: TV shows reality.
2-3	Antithesis: TV disturbs reality.
4-10	means: limited size of the screen result: curions effects
11-14	means: enlargement by means of zoomlens result: odd deceptions
15-16	assessment: irrelevant aspects to those who have become used to the fact that the camera does not lie.
17-24	distortion all the same dramatic effects in documentaries artistic effects (consciously achieved)

* Only those relationships are dealt with which are explicit in the text.

Jong, H.A.L. de	Cito Postbus 1034 6801 MG ARNHEM, The Netherlands
Kreeft, H. en P. Sanders	Cito Postbus 1034 6801 MG ARNHEM, The Netherlands
Lindblad, T.	The GUME/Tests Project Group University of Gothenburg, Department of Educational Research P.O. Box 1010 S431 26 MØLNDAL, Sweden
Maas, T.A.	Cito Postbus 1034 6801 MG ARNHEM, The Netherlands
Osborne, R.H.	Southern Regional Examinations Board Avondale House, 33 Carlton Crescent SOUTHAMPTON, SO9 4XL, England
Roosmalen, W.M. van	Cito Postbus 1034 6801 MG ARNHEM, The Netherlands
Samson, D.M.M.	Cito Postbus 1034 6801 MG ARNHEM, The Netherlands
Theunissen, P.J.J.M.	Cito Postbus 1034 6801 MG ARNHEM, The Netherlands
Wesdorp, H.	SCO Postbus 3753 1001 AW AMSTERDAM, The Netherlands

People who feel that they would be able to profit from the IUS meetings are invited to contact one of the co-chairmen and ask to be informed of current IUS activities and future meetings.

co-chairmen:

Christine Klein-Braley
Gesamthochschule Duisburg
Fachbereich 3 / Anglistik
Lotharstraße 65
D-4100 DUISBURG 1

Douglas K. Stevenson
Universität Essen
Fachbereich 3 / Anglistik
Universitätsstraße
D-4300 ESSEN 1